On top of the Higgs

Luca Colasurdo

Luca Colasurdo

On top of the Higgs

a measurement of the Higgs boson production
in association with top quarks

ON TOP OF THE HIGGS

A MEASUREMENT OF THE HIGGS BOSON PRODUCTION
IN ASSOCIATION WITH TOP QUARKS

LUCA COLASURDO

On top of the Higgs – a measurement of the Higgs boson production in association with top quarks
Thesis, Radboud University Nijmegen

ISBN: 978-94-028-1466-8

Cover design by Pina Saltarelli

Printing: Ipskamp Printing, Enschede

ON TOP OF THE HIGGS

A MEASUREMENT OF THE HIGGS BOSON PRODUCTION
IN ASSOCIATION WITH TOP QUARKS

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken
volgens besluit van het college van decanen
in het openbaar te verdedigen op dinsdag 21 mei 2019
om 16:30 uur precies

door

LUCA COLASURDO

geboren op 11 maart 1989
te Campobasso, Italië

PROMOTOREN	Prof. dr. N. de Groot Dr. F. Filthaut
MANUSCRIPTCOMMISSIE	Prof. dr. F.W.M. Verbunt Prof. dr. R.H.P. Kleiss Prof. dr. F. Linde (Universiteit van Amsterdam) Prof. dr. B. van Eijk (Universiteit van Twente) Dr. C.W.J.P. Timmermans

Contents

Introduction

With the expression *scientific method* we indicate the set of rules and procedures used to explore and investigate the world around us, to discover the fundamental laws of Nature. It starts with the observation of a phenomenon and the formulation of a hypothesis, a guess of what might be its explanation or its cause; the whole logic of scientific method is trying to disprove the original guess by performing an experiment whose results disagree with the predicted consequences of the original guess. If the guess is falsified by the experiment, we try to look for a better explanation; hence, in science, we are never sure we are right, we can only be sure we are wrong.

One of the most successful and well tested theories, without any doubt, is the Standard Model of Particle Physics. This theory is capable of explaining and describing, with a stunning level of precision, the interactions among matter's fundamental constituents.

In spite of being such a successful theory and having passed the test of time, one prediction was still not confirmed experimentally nor completely excluded until a few years ago: the existence of the Higgs boson, as the experimental proof of the correctness of the the Brout-Englert-Higgs mechanism for the origin of the masses of the elementary particles.

On the 4^{th} of July 2012, the ATLAS and CMS Collaborations announced the observation of a new particle consistent with the Higgs boson predicted by the Standard Model: it was its triumph.

This discovery was just the latest milestone of the long journey to understand how the world works. The journey began in Ancient Greece, in the moment the first philosophers rejected traditional mythological explanations in favour of more rational ones, based on direct observation of the phenomena and the capabilities of the human intellect. Democri-

tus was the first philosopher to postulate that the physical world consists of void, empty space filled with fundamental, indivisible building blocks: the atoms. This old idea remained alive through the centuries and, with the technological advancements of the 20$^{\text{th}}$ century, first the atoms and later the truly – so far – elementary particles have been discovered and included in the Standard Model.

The driving force of the search for the *first principles* of Nature was, and still is, our curiosity to understand the world around us: *"Considerate la vostra semenza: fatti non foste a viver come bruti, ma per seguir virtute e canoscenza"*[1] wrote Dante in the beginning of the fourteenth century in his Divine Comedy, to describe precisely this pursuit of knowledge.

However, this journey is far from ending, as the Standard Model cannot be the ultimate theory of Nature, given that it does not answer in a satisfactory way some open questions of fundamental importance: it does not incorporate the force of gravity into its description of the microscopic world; it describes only 5% of the matter and energy present in the whole universe, leaving out Dark Matter and Dark Energy; and it is not able to explain the amount of matter-antimatter asymmetry observed.

Furthermore, in spite of being more and more likely that the particle discovered is indeed the Higgs boson as predicted by the Standard Model, the possibility that it is simply one cousin in a family of many Higgses is still not ruled out completely. As a matter of fact, the existence of Beyond the Standard Model theories that can both answer these open questions and accommodate within themselves particles very similar to the discovered one makes it an interesting situation.

These theories predict the existence of a broad spectrum of new particles that have the potential to be produced and discovered at the LHC; however, there are no significant deviations from the Standard Model predictions until today. For this reason the interest in indirect searches has increased over the past years.

Precision measurements to test the internal consistency of the Stan-

[1] Dante Alighieri, "Divina Commedia", Inferno Canto XXVI, vv 118-120. Translation: "Consider how your souls were sown: you were not made to live like brutes or beasts, but to pursue virtue and knowledge".

dard Model are therefore needed, in particular to probe the effects of new particles in loop-induced processes. In this context, the precise determination of the so-called Yukawa coupling of the top quark to the Higgs boson is one of the key sectors, as most of Beyond the Standard Model theories predict a strong coupling to the last generation of quarks and leptons. As a consequence, the associated production of the Higgs boson with a pair of top quarks is one of the most promising areas of research to get the idea of the scale of new physics.

The renormalization evolution is a mathematical formalism that allows to describe the behaviour of coupling constants at different energy scales. Even if the Standard Model is a valid theory up to high energies, theoretical inconsistencies can appear as a consequence of the renormalization evolution of some coupling constants, i.e. when they become large, or the vacuum structure can change due to the development of additional minima in the Higgs potential.

The contribution of the top Yukawa coupling to the evolution of this potential at large values of the field itself is very important: a sub permill change in the coupling value can be responsible for the appearance of a second minimum, deeper than the one we are sitting in; depending on its precise value this means that our universe can be metastable and that the life-time of our vacuum is comparable to the age of the universe.

This thesis presents the work carried out within the context of the discovery and measurements of the properties of the Higgs boson. The work done by the Collaboration, to which I gave my little contribution, can be thought, as one more step along the road to discover and comprehend how Nature works.

Chapter 1 presents the theoretical background of the Standard Model, highlighting the relevant aspects for the studies presented in this thesis.

Chapter 2 describes the LHC and the ATLAS detector, as well as their performance.

Chapter 3 outlines the reconstruction and identification of the relevant physics objects with the ATLAS detector. I contributed to develop and validate part of the software framework used for the identification of b-jets that allows for the possibility to re-tag b-jets in the so-called *derivation* phase.

Chapter 4 describes the development of the the Jet Vertex Charge algorithm, as well as its calibration analysis. The first part presents the result of the work I did on its development, optimization and integration in the official ATLAS software, while the second part is dedicated to the calibration analysis of the algorithm.

The material presented in this chapter is entirely the results of the work I did in a team with my supervisors, which includes as well the post-doc who joined our team during my PhD.

Lastly, Chapter 5 contains the description of "one and a half" analyses, both on the search for the Standard Model Higgs boson produced in association with top quarks and decaying into a $b\bar{b}$ pair.

The "half" part refers to what will be referred to as "ICHEP analysis"; it presents my attempts to include the Jet Vertex Charge discriminant into the $t\bar{t}H(b\bar{b})$ analysis and the development of a method I worked on, named HFBDT, which aimed at improving the sensitivity of the analysis by improving the knowledge of the main irreducible background.

On the other hand, the "one" part presents the full search that contributed to first the evidence for and later the discovery of the $t\bar{t}H$ process. I contributed to the selection and the validation of the fit model for the profiled likelihood inputs to improve the analysis sensitivity.

Throughout the chapter, all the plots without any ATLAS label have been produced directly by me, while the plots carrying an ATLAS label come from public results to which I contributed.

The Standard Model and the Higgs boson 1

The Standard Model of Particle Physics (SM) is the theory that best summarizes the knowledge, as of today, of the subatomic world. It is able to describe with an unprecedented precision the interactions among all the fundamental particles: the *fermions*, the particles of matter; and the *bosons*, the force carriers. It describes three out of the four fundamental forces present in Nature: the electromagnetic, the weak and the strong force[1], based on the group symmetry $SU(3)_{\mathrm{C}} \otimes SU(2)_{\mathrm{L}} \otimes U(1)_{\mathrm{Y}}$, where the indices C, L and Y refer respectively to the quantum numbers of the colour, the chirality and the hypercharge of the particles.

Interactions are derived by the principle of *local gauge invariance* of this symmetry group. Finally, the spontaneous symmetry breaking mechanism is responsible for breaking the electroweak group into the electromagnetic one, leaving untouched the strong force group:

$$SU(3)_{\mathrm{C}} \otimes SU(2)_{\mathrm{L}} \otimes U(1)_{\mathrm{Y}} \rightarrow SU(3)_{\mathrm{C}} \otimes U(1)_{\mathrm{QED}} \tag{1.1}$$

This chapter is dedicated to a brief exposition of the Standard Model, with a particular emphasis on the Brout-Englert-Higgs mechanism.

1.1 The Standard Model as a gauge theory

The first steps towards the formulation of this model date back to the 1960s, in the attempt to unify of the electromagnetic and weak force by S. Glashow [1]. In 1967, S. Weinberg and A. Salam included the

[1] In this model is absent the fourth fundamental force, *gravity*, whose effects are too weak at the scales considered in this thesis. It is, in fact, hard to incorporate gravity, as described by General Relativity, in a consistent and coherent way in the framework of the Standard Model.

Brout-Englert-Higgs mechanism, responsible for the mass of the particles, in this theoretical framework, giving rise to the actual formulation of the electroweak theory [2, 3]. The strong interaction was included approximately at the same time, mainly by the work of R. Feynman, M. Gell-Mann and G. Zweig [4–6].

The framework used is the one of Quantum Field Theory (QFT) [7]: particles are treated as excitations of the fundamental field, as quanta of the field itself, and the dynamic variables of the theory are the quantized fields.

All known particles can be divided into two groups:

fermions are the constituents of matter and have half-integer spin. Depending on their properties they can be further divided into two groups: *leptons* and *quarks*, with the former not interacting via the strong force and the latter having colour charge and interacting with gluons. They are repeated into three generations, with the first generation being the lightest and most stable one and the others being heavier and unstable.

bosons are the force carriers and have integer spin. The photon is the mediator of the electromagnetic interaction; the W^+, W^- and Z mediate the weak interaction and gluons deal with the strong interaction.

Tables 1.1 and 1.2 summarize their most relevant properties, taken from Ref. [8].

The dynamics of the fields can be obtained using as a starting point the Lagrangian density, written in terms of the fields, from which it is possible to obtain the equation of motion through the least action principle. If the Lagrangian of the system is invariant under some transformations, it exhibits a symmetry and it is directly related to the conservation of some quantity in the system.

Symmetry is a fundamental concept in particle physics. From a mathematical point of view, whenever a transformation is applied to a set of equations and their solution does not change, then there is a symmetry involved. In case the parameter of the transformation is continuous, symmetries can be divided into space-time, as rotations and translations, and *internal* ones, when the transformation acts on the internal

Table 1.1: Summary of the electric charge and mass of the fermions in the SM, taken from Ref. [8]. Even if neutrinos are listed with zero mass, the recent evidence for neutrino oscillations indicates that they must have a non-zero mass [9, 10].

	Generation						
	I		II		III		charge [e]
	Flav.	Mass [MeV]	Flav.	Mass [MeV]	Flav.	Mass [GeV]	
Leptons	ν_e	0	ν_μ	0	ν_τ	0	0
	e	0.511	μ	105.66	τ	1.78	-1
Quarks	u	2.2	c	1.28 GeV	t	173.1	$+2/3$
	d	4.7	s	96	b	4.18	$-1/3$

Table 1.2: Summary of the electric charge and mass of the bosons in the SM.

Boson	Interaction	Charge	Spin	Mass [GeV]
γ	Electromagnetic	0	1	0
$W^\pm$	Weak	± 1	1	80.385 ± 0.015
Z		0	1	91.1876 ± 0.0021
Gluon	Strong	0	1	0

quantum numbers of the field. The latter group can be divided into *global* and local or *gauge symmetries*, depending on whether or not the transformation is different in every point of the space-time or not.

When a physical system has a continuous symmetry, Noether's theorem implies the existence of a conserved quantity [11]. As an example, from the invariance of the Lagrangian under translations it is possible to derive the law of momentum conservation.

The description of the interaction between particles arises in a natural way by requiring that the Lagrangian is invariant under a gauge transformation of a given symmetry group. Internal continuous symmetries lead to a conserved current and the introduction of a variable number of bosons, depending on the structure of the symmetry field.

Let us consider the invariance under a constant phase change of the

Lagrangian in the form of:

$$\begin{aligned} &\mathscr{L}(\psi', \overline{\psi}') = \mathscr{L}(\psi, \overline{\psi}) \\ &\text{if} \quad \psi \to \psi' = e^{ie\alpha}\psi \end{aligned} \tag{1.2}$$

From the field theory point of view, theory this effect is unnatural as the phase $e^{ie\alpha}$ is completely arbitrary, hence it does not contain any physics information. Besides, for a coherent definition of the field, every observer should agree upon the value of the phase, in contradiction to the principle that interactions are local and the principle of relativity, stating that there is no privileged observer. Therefore, a more natural choice is the use of a space-time dependent phase, $e^{ie\alpha(x)}$. In the language of group theory, it is a transformation under the abelian symmetry group $U(1)$.

Let us now consider the Dirac Lagrangian:

$$\mathscr{L} = \overline{\psi}(x)(i\gamma^\mu \partial_\mu - m)\psi(x) \tag{1.3}$$

where the symbol γ^μ represents the Dirac matrices. In order to preserve the *local* invariance it is necessary to add a new field, A_μ, which is able to reabsorb the extra term given by the derivative acting on the phase. For this purpose, the introduction of the *covariant derivative* is needed:

$$\partial_\mu \to D_\mu = \partial_\mu - ieA_\mu \tag{1.4}$$

As a consequence, the fields now transform as:

$$\begin{aligned} &\psi(x) \to \psi(x)' = e^{ie\alpha(x)}\psi(x) \\ &A_\mu(x) \to A_\mu(x)' = A_\mu(x) + \partial_\mu \alpha(x) \end{aligned} \tag{1.5}$$

The additional field A_μ is identified with the photon field and, in order to obtain a complete dynamic theory, it is necessary to add a kinetic term for the photon. In this way, the Lagrangian of the Quantum ElectroDynamics (QED) is derived as

$$\mathscr{L}_{\text{QED}} = \overline{\psi}(i\gamma^\mu \partial_\mu + e\gamma^\mu A_\mu - m)\psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} \tag{1.6}$$

The term $A_\mu J^\mu = eA_\mu \overline{\psi}\gamma^\mu \psi$ represents the interaction between the photon and the electron, while the term $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ describes the motion of the photon. This interaction can be schematically represented using the so-called Feynman diagrams, as shown in Figure 1.1. Such pictorial representation is an important tool used in perturbation theory to compute the probability of a process up to a specified perturbation order.

If time flows from left to right and the space-axis is vertical, this diagram can be read as a photon, the wavy line, creating an electron-positron pair, the solid line; the arrow is indicative of the flow of the electric charge, therefore a positron (or an antiparticle in general) can be thought as an electron (particle) moving backwards in time. The strength of the interaction is proportional to the electromagnetic charge.

It is worth stressing that the complete form of the interaction is derived based uniquely on the symmetries of the Lagrangian for the free electron field, Eq. (1.3), and that the full QED Lagrangian is able to predict with a remarkable precision a great variety of phenomena[2].

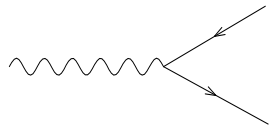


Figure 1.1: Feynman diagram of the electron-photon interaction ($eA_\mu \overline{\psi}\gamma^\mu \psi$). Considering the time-axis oriented from left to right and space-axis in the vertical direction, the wavy line represents a photon and the solid lines the creation of an electron-positron pair.

1.1.1 The strong sector: Quantum ChromoDynamics

The sector of the Standard Model responsible for the strong interaction is based on the non-abelian group $SU(3)_C$, where the subscript refers to

[2] For example, the anomalous magnetic moment of the electron, $g-2$, was experimentally determined with unprecedented precision in 2008 [12] and compared to an independent measurement of the fine structure constant, α, using computations including eight-order QED contribution: the agreement of the two measurements is the most stringent test of the QED performed so far, at a level better than 10^{-9} [13].

the colour charge of the quarks, in an analogous way of QED.

The derivation of the QCD Lagrangian is a simple exercise of gauge theory. The starting point is the Dirac Lagrangian for a quark:

$$\mathscr{L} = \overline{\mathbf{q}}(i\gamma^\mu \partial_\mu - m)\mathbf{q} \qquad \mathbf{q} = \begin{pmatrix} q_1 \\ q_2 \\ q_3 \end{pmatrix} \tag{1.7}$$

where the indices $i = 1,2,3$ refer to the colour of the quark.

The colour symmetry is realized with:

$$\mathbf{q} \to U\mathbf{q} \qquad \overline{\mathbf{q}} \to \overline{\mathbf{q}}U^\dagger \tag{1.8}$$

with $U \in SU(3)$. It is possible to express the matrix in terms of the generators of the group, as $U = \exp\{i\frac{\lambda_a}{2}\theta^a\}$, where the λ_a matrices are the Gell-Mann matrices and identify a base of the $SU(3)$ algebra, with the index a running over the eight generators of the group[3].

The covariant derivative is obtained using the same procedure as for the case of QED:

$$\begin{aligned} D_\mu q &= (\partial_\mu + ig_s G_\mu)\, q \\ [G^\mu]_{\alpha\beta} &\equiv \frac{1}{2}\lambda^a_{\alpha\beta} G^\mu_a \end{aligned} \tag{1.9}$$

so that eight gluon fields are introduced, G^a_μ, one for each generator of the group. The indices $\alpha, \beta = 1,2,3$ are the row and column indices of the λ_a matrices, which represent the colour of the quark.

Following the Yang-Mills prescription [14], which is nothing but a generalization of the QED process to a non-abelian case, it is straightforward to write down the kinetic term for the gluon field, so that the QCD Lagrangian is found to be:

$$\mathscr{L}_{\text{QCD}} = \sum_q \overline{\psi}_{q,\alpha}(i\gamma^\mu \partial_\mu \delta_{\alpha\beta} - g_s \gamma^\mu \tfrac{\lambda^a_{\alpha\beta}}{2} G^a_\mu - m_q \delta_{\alpha\beta})\psi_{q,\beta} - \tfrac{1}{4} G^a_{\mu\nu} G^{a,\mu\nu} \tag{1.10}$$

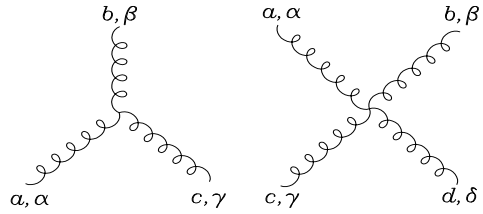
[3] In a generic group $SU(N)$ there are $N^2 - 1$ generators, i.e. matrices that form a basis of the said group algebra.

The field $\psi_{q,\alpha}$ is the spinor for the quark field q with colour α and the field G^a_μ represents the a-th gluon, g_s is the coupling constant of the strong force. Finally the tensor $G^a_{\mu\nu}$ is given by:

$$G^a_{\mu\nu} = \partial_\mu G^a_\nu - \partial_\nu G^a_\mu - g_s f_{abc} G^b_\mu G^c_\nu, \qquad [\lambda^a, \lambda^b] = 2i f_{abc} \lambda^c \quad (1.11)$$

where the f_{abc} are called the *structure constants* of the group and for the $SU(3)$ group are fully antisymmetric in the exchange of any two indices and vanish in case two of the indices are equal.

The non-abelian nature of the group marks the biggest difference with respect to the QED case; in fact, this is responsible for the third term of $G^a_{\mu\nu}$ in Eq. (1.11) which leads to an interaction among the force mediators themselves with the same coupling constant g_s. The resulting tri-linear and quadri-linear gluon self-interaction are represented with the Feynman diagrams in Figure 1.2.



(a) tri-linear vertex **(b)** quadri-linear vertex

Figure 1.2: Feynman diagrams showing the tri-linear and quadri-linear gluon self-interaction.

Finally, the gluon self-interaction has an effect on the way the strong coupling evolves as a function of the energy (distance): it becomes weaker for higher energies (shorter distances), a property known as *asymptotic freedom*, and is stronger for lower energies, *confining* quarks into hadrons. This feature marks the difference with respect to the electromagnetic force, as the QED coupling strength increases as the energy increases.

1.1.2 The electroweak sector

The electroweak sector of the Standard Model of Particle Physics is based upon the symmetry group $SU(2)_{\mathrm{L}} \otimes U(1)_{\mathrm{Y}}$, where the group $SU(2)_{\mathrm{L}}$ refers to the *weak isospin* charge, I, while $U(1)_{\mathrm{Y}}$ refers to the *weak hypercharge* Y. The left-handed components, i.e. the components with *chirality* left (L), are organized in doublets with $I = 1/2$, while the right-handed components (R) in singlets. Chirality is the phenomenon for which the mirror image and the original do not behave in the same way. It plays an important role within the SM, as the right and left components are treated differently by the weak interactions: only left left-handed fermions and right-handed antifermions interact with the $W^{\pm}$ bosons.

Table 1.3 summarizes the quantum numbers of the fermions related to the electroweak sector of the SM.

Table 1.3: Summary of the quantum numbers of the fermions in the SM. The subscripts L and R refer to the chirality states left and right respectively. The superscript for the quarks indicates that they are eigenstates of the electroweak interactions, hence superposition of the mass eigenstates, for which the mixing is described by the CKM matrix, as explained in Section 1.1.4. Neutrinos with right-handed chirality are not included in the SM.

	Generation			Quantum numbers			
	1	2	3	I	I_3	Y	Q [e]
Leptons	$\begin{pmatrix} \nu_e \\ e^- \end{pmatrix}_{\mathrm{L}}$	$\begin{pmatrix} \nu_\mu \\ \mu^- \end{pmatrix}_{\mathrm{L}}$	$\begin{pmatrix} \nu_\tau \\ \tau^- \end{pmatrix}_{\mathrm{L}}$	$1/2$ $1/2$	$1/2$ $-1/2$	-1 -1	0 -1
	e^-_{R}	μ^-_{R}	τ^-_{R}	0	0	-2	-1
Quarks	$\begin{pmatrix} u \\ d' \end{pmatrix}_{\mathrm{L}}$	$\begin{pmatrix} c \\ s' \end{pmatrix}_{\mathrm{L}}$	$\begin{pmatrix} t \\ b' \end{pmatrix}_{\mathrm{L}}$	$1/2$ $1/2$	$1/2$ $-1/2$	$1/3$ $1/3$	$2/3$ $-1/3$
	u_{R}	c_{R}	t_{R}	0	0	$4/3$	$2/3$
	d_{R}	s_{R}	b_{R}	0	0	$-2/3$	$1/3$

In order to obtain the interaction Lagrangian, the procedure is the

same as the one used in the previous sections: one first imposes the $SU(2)$ and $U(1)$ symmetries, then applies the gauge principle and introduces a covariant derivative that takes the form of:

$$D^{\mu} = \partial^{\mu} + ig\frac{\sigma_i}{2}W_i^{\mu} + ig'\frac{Y}{2}B^{\mu} \tag{1.12}$$

where g and g' are the coupling constants of the gauge boson associated with the $SU(2)_{\mathrm{L}}$ and $U(1)_{\mathrm{Y}}$ symmetries respectively, σ_i are the Pauli matrices and Y is the hypercharge.

The kinetic part of the Lagrangian takes the form:

$$\mathscr{L}_{\mathrm{EW}} = i\overline{\psi}(\gamma^{\mu}D_{\mu})\psi - \frac{1}{4}W^i_{\mu\nu}W_i^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} \tag{1.13}$$

After imposing these symmetries, we are left with four vector bosons: three coming from the $SU(2)$ group, the fields $W_{\mu}^{1,2,3}$, and one vector boson for the $U(1)$ group, B_{μ}. The physical fields are obtained by a linear combination of them:

$$W_{\mu}^{\pm} = \frac{W_{\mu}^1 \mp iW_{\mu}^2}{\sqrt{2}} \tag{1.14}$$

and the Z boson and the photon are a linear combination of the two remaining vector bosons, W_{μ}^3 and B_{μ}:

$$\begin{aligned} A_{\mu} &= \sin\theta_{\mathrm{w}}W_{\mu}^3 + \cos\theta_{\mathrm{w}}B_{\mu} \\ Z_{\mu} &= \cos\theta_{\mathrm{w}}W_{\mu}^3 - \sin\theta_{\mathrm{w}}B_{\mu} \end{aligned} \tag{1.15}$$

The $W^{\pm}$ bosons are mediators of the charged currents, A mediates the electromagnetic current and the Z boson is the neutral current mediator. The parameter θ_{w} is called *Weinberg's angle*.

The photon interacts with all charged particles with the coupling being the electric charge and does not distinguish between right- and left-handed fermions. The relation between the weak isospin, the hypercharge and the electric charge Q is given by:

$$Q = I_3 + \frac{Y}{2} \tag{1.16}$$

and the relation among the couplings and the electric charge is:

$$e = g \sin\theta_{\mathrm{w}} = g' \cos\theta_{\mathrm{w}} \tag{1.17}$$

It should have been noted in the previous cases that there is no mass term for the gauge boson in the Lagrangian [4]. Following Glashow's example, it is possible to add "by hand" a mass term to it, whose effect is to break the gauge symmetry used to build the Lagrangian. This addition makes the theory *non renormalizable*[5], therefore an alternative is needed to give the mass to the different particles, as it is possible to see in the next section.

1.1.3 The Brout-Englert-Higgs mechanism

The theory exposed up to now predicts that all gauge bosons have a null mass, like the photon; this is clearly false for the vector gauge bosons $W^{\pm}$ and Z, whose mass is respectively 80.385 ± 0.015 GeV and 91.1876 ± 0.0021 GeV [8]. Adding to the Lagrangian a mass term by hand will result in breaking the gauge invariance and will make the theory non-renormalizable, therefore a different method to dynamically add a mass term needs to be found.

The problem has been solved using the concept of *Spontaneous Symmetry Breaking* (SSB). This concept was first introduced by Nambu and Goldstone [15–17], predicting the appearance of a number of massless scalar bosons. Later in 1964 several papers successfully introduced spontaneously-broken local symmetry into the model of the electroweak interactions, first by Englert and Brout [18] and followed a few weeks later by two papers written by Higgs [19, 20], who did not know about their work. Later that year, an article signed by Guralnik, Hagen and Kibble [21] developed the same ideas on symmetry breaking with gauge invariance[6].

[4] Also a fermion mass term is missing in the Lagrangian, as it would couple the both the right- and left-handed chirality particles and would break the gauge invariance.

[5] A theory has to be renormalizable in order to provide meaningful predictions. Renormalization is the procedure that allows to absorb divergences that arise during computations in perturbative calculations beyond leading order and hide them into a redefinition of the charge, mass and fields of the various particles.

[6] Despite there are references to the work of the other physicists, Guralnik stated that their

The importance of the Higgs paper is that he was the only author to explicitly predict the existence of a massive scalar boson. Therefore, it is often called *Higgs mechanism* for simplicity.

The simplest and most elegant way to break a symmetry and let the various particles acquire mass is via the introduction of an extra scalar field, whose ground state is not invariant under the original symmetries present in the Lagrangian.

In order to spontaneously break the symmetry of the group $SU(2)_{\mathrm{L}} \otimes U(1)_{\mathrm{Y}}$ there is the need to introduce a scalar field that is an isospin doublet:

$$\phi = \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix}_{Y=1} \tag{1.18}$$

where the fields ϕ_1 and ϕ_2 are complex fields.

The simplest Lagrangian takes the form:

$$\begin{aligned} \mathscr{L}_{\mathrm{Higgs}} &= (D_\mu \phi)^\dagger (D^\mu \phi) - V(\phi) \\ &= (D_\mu \phi)^\dagger (D^\mu \phi) - \mu^2 \phi^\dagger \phi - \lambda (\phi^\dagger \phi)^2 \end{aligned} \tag{1.19}$$

and the covariant derivative is the one in Eq. (1.12)

The potential $V(\phi)$ depends on two parameters: μ^2 and λ. The condition $\lambda > 0$ is sufficient to ensure that the spectrum of the energy levels has a lower bound. If the parameter μ^2 is positive, the potential has only one minimum for $\phi = 0$.

The more interesting case happens when $\mu^2 < 0$ and the symmetry of the potential is broken. The shape of the potential is shown in Figure 1.3. In reference to this figure, the ground state is not in $\phi = 0$ anymore, but lies on a circumference for which $|\phi|^2 = \text{const}$. The evolution of the system will spontaneously move it in its ground state and choosing one of them will spontaneously break the symmetry of the initial condition of the system. The value of the field in the ground state is called *vacuum expectation value* (VEV), v, and it is related to the

work "was done in its entirety without any knowledge of others working on the same problem of symmetry breaking and gauge system" [22].

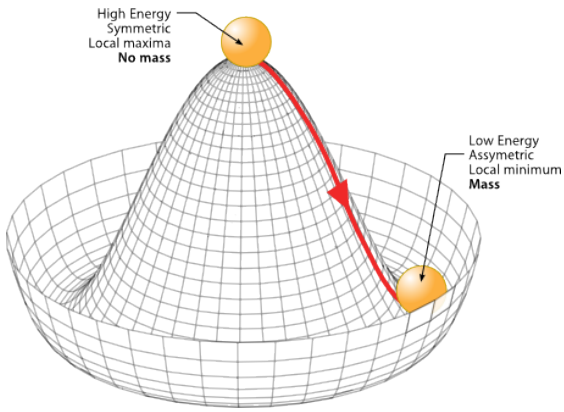


Figure 1.3: The Higgs potential for the case $\lambda > 0$ and $\mu^2 < 0$.

parameters in Eq. (1.19) via the relation:

$$\phi_0^2 = -\frac{\mu^2}{2\lambda} \equiv \frac{v^2}{2} \tag{1.20}$$

In order to obtain a description in terms of particles, that is fluctuations or quanta of the field, it is necessary to expand the field around a stable state, which translates into choosing one of the degenerate ground states in a unique way.

The field can be parametrized as:

$$\phi(x) = \frac{1}{\sqrt{2}} e^{i\sigma_i \theta_i(x)/v} \begin{pmatrix} 0 \\ v + H(x) \end{pmatrix} \tag{1.21}$$

such that the real field $H(x)$ represents the radial fluctuations around the equilibrium and the real fields $\theta_i(x)$ represent the angular excitations that leave the energy of the system unchanged.

It is possible to perform a gauge transformation to remove the non-physical degrees of freedom, which will make the $\theta(x)_i$ fields disappear, so that the only physical field left is $H(x)$:

$$\phi(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + H(x) \end{pmatrix} \tag{1.22}$$

Substituting the parametrization in Eq. (1.22) into the equation of

the covariant derivative, one obtains terms proportional to $W_\mu^+ W^{-\mu}$ and $Z_\mu Z^\mu$, which represent mass terms for the gauge boson[7]. In this way, the $W^\pm$ and Z bosons acquire masses equal to:

$$m_W = \frac{1}{2} g v, \qquad m_Z = \frac{1}{2} v \sqrt{g^2 + g'^2} \tag{1.23}$$

while the photon remains massless and the Higgs boson acquires a mass of:

$$m_H = \sqrt{\lambda v^2} \tag{1.24}$$

It should be noted that the mass of the Higgs boson is not predicted by the model, given its dependence on λ, but once it is fixed all other properties can be computed theoretically.

Furthermore, this model predicts a relation between the masses of the W and Z boson, named *custodial symmetry*:

$$\rho \equiv \frac{m_Z^2 \cos^2 \theta_\mathrm{w}}{m_W^2} = 1 \tag{1.25}$$

which is true at tree level in perturbation theory and radiative corrections give a few percent deviation[8] [23].

1.1.4 Fermion masses

In the formulation of the SM, the fermion mass terms cannot be included in the Lagrangian, as they will couple right- and left-handed particles, whereas the $SU(2)$ symmetry treats those two as different species.

Luckily, it is possible to give mass to fermions by introducing a so-called *Yukawa interaction* between a left-handed fermion doublet, the Higgs field and a right-handed fermion singlet.

[7] It is possible to interpret this as saying that three of the four real scalar fields have been *absorbed* or *eaten* by the vector fields to acquire mass.

[8] Thanks to the computation of such radiative corrections and a precise measurement of those parameters, it was possible to predict the top quark mass with good precision before its discovery.

In case of leptons, considering just one family, the Yukawa term is given by the following equation:

$$\mathscr{L}_{\text{Yukawa}} = -g_f \overline{L}_{\text{L}} \phi l_{\text{R}} + \text{h.c.} \quad \text{with} \quad \overline{L}_{\text{L}} = \begin{pmatrix} \nu_{\text{L}} \\ l_{\text{L}} \end{pmatrix} \tag{1.26}$$

where $\overline{L}_{\text{L}}$ and l_{R} represent the lepton doublet and singlet respectively and h.c. stands for the Hermitian conjugate. After the substitution of the value of ϕ given by Eq. (1.22), it becomes:

$$\mathscr{L}_{\text{Yukawa}} = -\frac{g_f \nu}{\sqrt{2}} \overline{\psi}\psi - \frac{g_f}{\sqrt{2}} \overline{\psi}\psi H \tag{1.27}$$

where the first term represents the lepton mass, which is proportional to the Higgs VEV, while the second term is the interaction between the lepton and the Higgs boson, with a strength given by the same coupling parameter g_f.

The case of quarks is more complex. The Yukawa interaction terms can be written as:

$$\begin{aligned} \mathscr{L}_{ij} &= -g^D_{ij} \overline{Q}_i \phi D_j - g^U_{ij} \overline{Q}^a_i \varepsilon_{ab} \phi^{*b} U_j + \text{h.c.} \\ \mathscr{L}_{\text{qH}} &= \sum_{ij} \mathscr{L}_{ij} \end{aligned} \tag{1.28}$$

where the indices $i, j = 1,2,3$ are the index over the families, the indices $a, b = 1,2$ are the weak isospin and are summed, and ε is the antisymmetric tensor. Q_i is a generic left-handed doublet, while U_i and D_i denote the right-handed fields of up (= u, c, t) and down (= d, s, b) type:

$$Q_i = \begin{pmatrix} u \\ d \end{pmatrix}_{\text{L}}, \begin{pmatrix} c \\ s \end{pmatrix}_{\text{L}}, \begin{pmatrix} t \\ b \end{pmatrix}_{\text{L}} \tag{1.29}$$

$$U_i = u_{\text{R}}, c_{\text{R}}, t_{\text{R}} \qquad D_i = d_{\text{R}}, s_{\text{R}}, b_{\text{R}} \tag{1.30}$$

expressed as eigenstates the weak interaction.

By substituting Eq. (1.22), the part of the Lagrangian relative to the a mass term for the quarks becomes:

$$\begin{aligned} \mathscr{L}_{\text{qm}} &= \overline{D}_{\text{L}} M^{\text{d}} D_{\text{R}} + \overline{U}_{\text{L}} M^{\text{u}} U_{\text{R}} + \text{h.c.} \\ M^{\text{d}}_{ij} &= \frac{g^D_{ij} v}{\sqrt{2}}; \qquad M^{\text{u}}_{ij} = \frac{g^U_{ij} v}{\sqrt{2}} \end{aligned} \tag{1.31}$$

and the $M^{\text{d,u}}$ matrices are, in general, non diagonal.

In the preceding expressions, the quarks are expressed as eigenstates of the weak interaction. The mass eigenstates are obtained by diagonalizing the matrices $M^{\text{u,d}}$ as:

$$M^{\text{u,d}}_{\text{diag}} = V^{\text{u,d}}_{\text{L}} M^{\text{u,d}} {V^{\text{u,d}}_{\text{R}}}^{\dagger} \tag{1.32}$$

It is possible to incorporate the unitary complex matrices V into a redefinition of the fields of the quarks. As a result, the $W^{\pm}$ interactions in the Lagrangian, after going from the interaction to the mass eigenstates, have a coupling that depends on the quarks involved in the interaction; more precisely, they include an element of the so-called Cabibbo, Kobayashi and Maskawa matrix [24, 25], V_{CKM}, defined as:

$$V_{\text{CKM}} = {V^{\text{u}}_{\text{L}}}^{\dagger} V^{\text{d}}_{\text{L}} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \tag{1.33}$$

The elements of the V_{CKM} give the intensity of the Charged Current, which permits the change of quark flavour. They explain the different rates of decay into leptons observed experimentally, thus preserving what is referred to as *lepton universality*, i.e. the coupling of the vector bosons to leptons is the same, regardless of the flavour of the lepton itself.

This matrix is an extension of the Cabibbo theory; as a matter of fact, the elements V_{ud} and V_{us} are equal to $\cos\theta_{\text{c}}$ and $\sin\theta_{\text{c}}$, where θ_{c} is the Cabibbo angle, as can easily be seen using Wolfenstein's parametrization [26]. Elements on the diagonal are of order 1, whereas off-diagonal elements are suppressed, disfavouring weak transitions between different families.

A final note on the CKM matrix is that if there are at least three quark families there is at least one irreducible phase in the CKM matrix, which is the only source of CP violation in the SM.

As for the lepton case, after diagonalizing the mass matrix, the quark mass is proportional to the Higgs VEV via the same coupling, g_f, that controls the strength of the interaction between the Higgs boson and the other fermions:

$$m_f = \frac{v}{\sqrt{2}} g_f \tag{1.34}$$

In summary, the Standard Model is a renormalizable quantum field theory that is able to predict with extraordinary precision particle interactions by the means of perturbation theory. The free parameters of the theory are:

- 3 coupling constants: g, g' and g_s for the groups $U(1)_{\mathrm{Y}}$, $SU(2)_{\mathrm{L}}$ and $SU(3)_{\mathrm{C}}$ respectively;
- 2 parameters for the electroweak symmetry breaking mechanism: v and m_H;
- 9 Yukawa couplings for the fermion masses;
- 4 parameters for the CKM matrix.

1.2 Phenomenology of the Higgs at the LHC

In this section the production mechanism and the main features of the Higgs boson will be illustrated.

The LHC is a proton-proton collider and one of its main goals was the discovery of the Higgs boson – or provide conclusive evidence of its absence. The production of new particles arises from the interaction between quarks and gluons, therefore it is important to know the physics of pp collisions.

1.2.1 Proton-proton interactions

In a pp machine, the colliding particles are composite particles. In the parton model, hadrons are seen as made up with a collection of quarks and gluons; in fact, other than the three valence quark, in reality there is much more going on: virtual pairs of quark anti-quark are continuously created and annihilated by quantum fluctuations and gluons are binding all of them together.

Based on the energy, it is possible to divide such interactions into perturbative QCD, with a high momentum transfer, and non-perturbative QCD. In the former case, it is possible to *factorize* the process into the hard scattering and the extraction of the parton from the proton. That means that a pp collision has to be seen as an interaction among its constituents, given the centre of mass energy of the collisions.

In Figure 1.4a is depicted a schematic diagram of a hadron hadron scatter. In mathematical terms it can be written down as in the following equation:

$$\sigma(s,\mu_{\mathrm{F}}) = \sum_{a,b} \int_0^1 f_a(x_1,\mu_{\mathrm{F}}) f_b(x_2,\mu_{\mathrm{F}}) \hat{\sigma}(\hat{s},\mu_{\mathrm{F}})\, \mathrm{d}x_1 \mathrm{d}x_2 \tag{1.35}$$

where a, b are the two partons involved in the process, x_i is the fraction of the proton momentum carried by the parton i and $f_i(x)$ represents the density of partons (of type i) in the proton to carry a fraction x of the proton momentum, named *Parton Density Function* (PDF). Finally, $\hat{\sigma}(\hat{s})$ is the cross-section of the hard scatter computed using perturbative QCD and happening with a squared centre of mass energy $\hat{s} = x_1 x_2 s$.

The factorization scale, μ_{F}, can be seen as the resolution at which the hadron is being probed, the scale at which is it possible separate hard scattering processes from the non-perturbative ones.

Because of the large energy transfer, the protons will usually break up: the remnants of the protons will form what is called *underlying event*.

As discussed at the end of Section 1.1.1, QCD has the property of colour confinement at large distances. This translates into the fact that quarks cannot be directly observed as free particles, resulting in a very complicated final state, shown by the the tree-like structure depicted in

Figure 1.4b. Final state quarks or gluons emit softer gluons and gluons can split into a quark/anti-quark pair, a process called *showering*, until the showering reaches an end and colour-neutral hadrons are formed, a process called *hadronization*. The showering process is illustrated by the red and purple gluon emissions and the hadronization by the green blobs in the same figure.

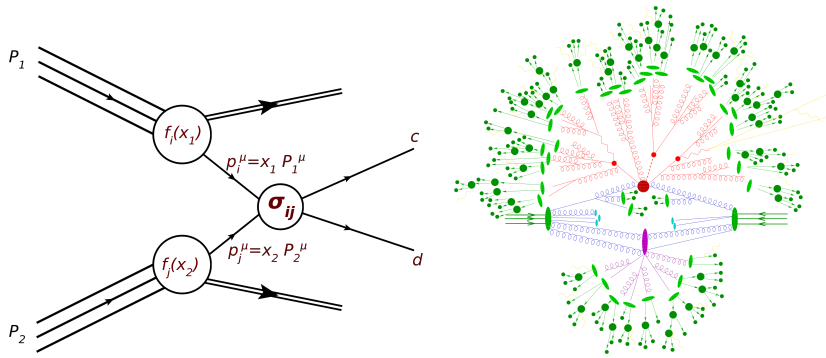


Figure 1.4: (a) A schematic diagram of hadron-hadron collision which shows the hard scattering process for the production of final states c and d. (b) Sketch of a hadron-hadron collision as simulated by a Monte Carlo event generator. The red blob in the centre represents the hard collision, surrounded by a tree-like structure representing the showering and hadronization part of the process.

QCD does not predict the PDF, hence their shape is extracted by a fit to data from a variety of different sources, such as deep inelastic scattering, jet p_{T} or rapidity spectra, as well as vector boson rapidity and transverse momentum distributions. Once they are know at an energy scale, they are subsequently extrapolated to different energies using DGLAP evolution equations [27–30].

1.2.2 Higgs boson production modes

Given the centre of mass energy of the collisions, 7 and 8 TeV in Run1 and 13 TeV in Run2, the LHC can largely be seen as a gluon collider,

since the corresponding PDFs imply that it is more likely to pick up low-x gluons from the proton. The main Higgs boson production mechanisms are:

gluon-gluon fusion (ggF) is the main production mode at the LHC, even if the Higgs boson does not couple directly to gluons. The Feynman diagram responsible for this process is shown in Figure 1.5a. Even if in principle, the Higgs boson couples to all massive particles, with a coupling proportional to their mass, in practice, the main contributions in the loop are coming from the top and bottom quarks.

Vector Boson Fusion (VBF) is the process in which the initial quarks radiate two virtual W or Z bosons that later annihilate to produce a Higgs boson. The peculiarity of this process is the presence of two energetic quarks in the final state emitted mainly in the forward and backward regions of the detector, whereas the Higgs boson decays in the central region. Furthermore, given that it is a purely electroweak process, the hadronic activity in the centre of the detector is very low. The Feynman diagram corresponding to this process is shown in Figure 1.5b.

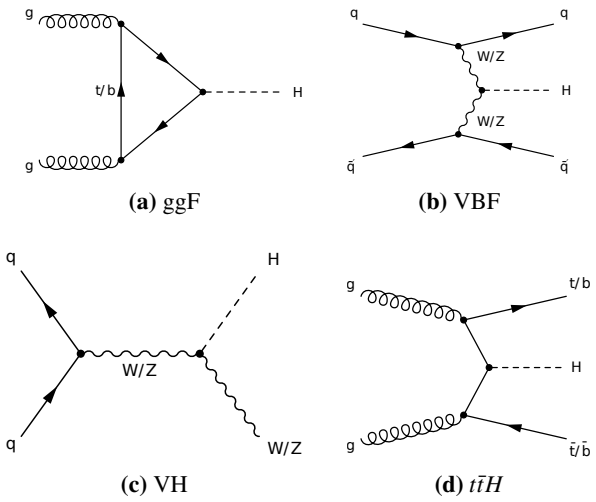
Associated production with a vector boson (VH) or *Higgsstrahlung* has its Feynman diagram shown in Figure 1.5c. The presence of a W or Z boson in the final state is often used to identify the events and suppress the backgrounds.

Associated production with a pair of top quarks ($t\bar{t}H$) is the Higgs production mode on which this thesis will focus. It has the lowest cross-section compared to the previous modes, but on the other hand, it is the only one that can be used to measure in a direct way the coupling of the Higgs to the top quark. The cross-section is equal to 507^{+35}_{-50} fb, as calculated at Next-to-Leading-Order (NLO) in QCD, for a Higgs boson of 125 GeV of mass at $\sqrt{s}$ = 13 TeV. The corresponding Feynman diagram is shown in Figure 1.5d[9].

[9] This is actually just one of the Feynman diagrams for this process, as the Higgs boson could be radiated off of one of the two top quarks as well.

Figure 1.6 shows the predicted production cross section as a function of the centre of mass energy; at $\sqrt{s}$ = 13 TeV the total cross section for a Higgs boson of mass m_H = 125.09 GeV is:

$$\sigma(pp \to H) = 50.43\ ^{+10,1\%}_{-12,9\%}\ (\text{scale})\ ^{+12,1\%}_{-11,5\%}\ (\text{PDF} + \alpha_s)\ \text{pb} \qquad (1.36)$$



(a) ggF **(b)** VBF

(c) VH **(d)** $t\bar{t}H$

Figure 1.5: Feynman diagrams corresponding to the main Higgs production modes at the LHC.

1.2.3 Higgs boson branching ratios

As reported in the previous sections, the Higgs boson can decay into a pair of fermions and bosons with a coupling proportional to the particle's mass, i.e. the heavier the particle, the more likely the decay is, provided that the pair of particles it decays to is light enough so that it can be produced on shell.

All the values for the Branching Ratios (BR) presented in this subsection are taken from the Handbook of LHC Higgs Cross Sections [32].

The most important decay is into a $b\bar{b}$ pair, which occurs in about 57.7% of the cases and will also be the main focus of this thesis. Con-

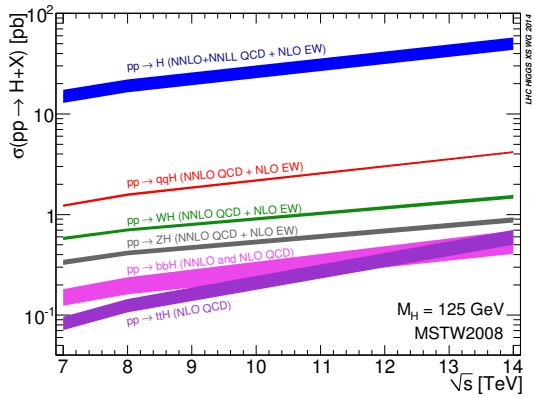


Figure 1.6: Higgs boson cross-section as a function of the centre of mass energy. Taken from Ref. [31].

tinuing with the fermions, comes the decay channel into a $\tau^+\tau^-$ pair, which accounts for about 6% of the BR. In spite of this, it is important in order to test the coupling with leptons.

Looking at decays into bosons, the decay into WW pair is the most important (21.5%), but given the low mass of the Higgs boson, one of the bosons has to be produced off-shell. Both the ZZ (2.64 %) and $\gamma\gamma$ (0.22%) channels have a very low BR but offer the possibility to fully reconstruct the final state and the invariant mass of the Higgs boson with an excellent resolution, allowing for a strong suppression of the backgrounds and a high analysis sensitivity.

For m_H = 125 GeV, the width is $\Gamma_H = 4.07 \pm 0.16$ MeV, below the experimental resolution.

1.2.4 The discovery of the Higgs boson

The great success of the Standard Model derives not only from its simple description of three out of four fundamental forces, but also from its very high predictive power and its extreme precision.

After the discovery of the $W^\pm$ and Z bosons [33, 34] and the top quark [35, 36], there has been a tremendous effort in order to find the last missing piece.

In the framework of the SM, physics observables are computed using perturbation theory; higher order corrections can therefore be sensitive to particles produced in the loop diagrams. Precision measurements obtained at LEP, SLC and Tevatron [37–39], combined with high accuracy in the theory calculations, allowed to test the internal consistency of the SM and put indirect constraints on properties of unknown particles. Figure 1.7 shows the result of a χ^2 fit, done early in 2012, using as inputs precision measurements of parameters of the electroweak sector sensitive to the Higgs mass, as a function of the putative mass of the Higgs boson; grey regions are excluded by direct searches performed at LEP and Tevatron.

On July 4th 2012, at CERN, the ATLAS and CMS Collaborations announced the observation of a new particle compatible with the Higgs boson as predicted by the SM [40, 41]. The discovery was mainly driven by analyses searching for the Higgs decaying into bosons. Figure 1.8 shows the *p-value*[10], p_0, for the combination of the individual searches for the Higgs boson decaying into a pair of photon, four leptons and $H \to WW^* \to e\nu\mu\nu$ at the end of July of the same year[11].

1.2.5 The Run1 result

In 2015 the ATLAS and CMS Collaborations released the first combined measurement of the Higgs boson mass [42], obtained from a simultaneous fit to the reconstructed invariant mass peaks in the $H \to \gamma\gamma$ and $H \to ZZ \to 4l$ channels:

$$m_H = 125.09 \pm 0.21(\text{stat}) \pm 0.11(\text{syst}) \text{ GeV} \tag{1.37}$$

Furthermore, both collaborations released a combined measurement of the Higgs boson production and decay rates, as well as constraints on its couplings to vector bosons and fermions [43]. The combination is

[10] The p-value is a measure of how luckily it is to observe a fluctuation in the backgrounds at least as extreme as the one observed in data. Low values indicate big discrepancy between the observed data and the hypothesis that the Higgs boson is not present.

[11] For the sake of correctness, it should be made clear that this is not the plot shown during the CERN seminar announcing the discovery, as it includes also the analysis searching for the decay $H \to WW^* \to e\nu\mu\nu$ using 2012 data, whereas the results shown during the seminar announcing the discovery were using only the data collected during 2011 for this particular analysis.

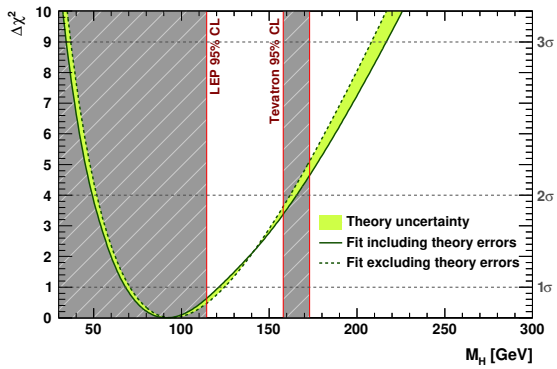


Figure 1.7: Indirect determination of the Higgs boson mass: $\Delta\chi^2$ as a function of m_H. Grey bands represent regions excluded by direct searches. Taken from Ref. [37].

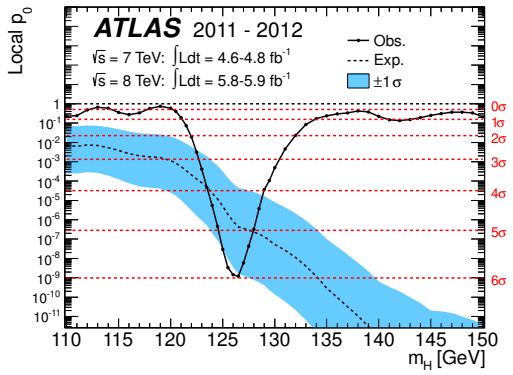
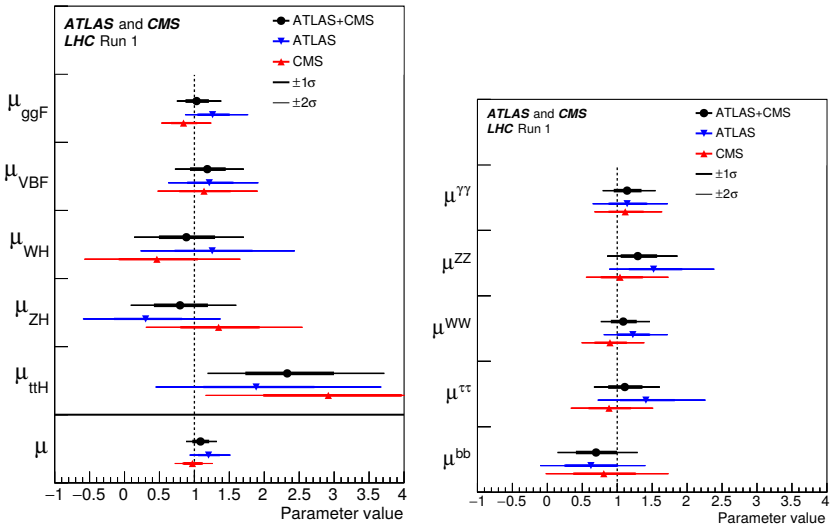


Figure 1.8: The observed (solid line) local p_0 as a function of m_H in the low mass range. The dashed curve shows the expected local p_0 under the hypothesis of a SM Higgs boson signal at that mass with its $\pm 1\sigma$ band. The horizontal dashed lines indicate the p-values corresponding to significances of 1 to 6 σ. The largest observed local significance was found for a SM Higgs boson mass hypothesis of m_H = 126.5 GeV, where it reached 6.0 σ. Taken from Ref. [40].



(a) **(b)**

Figure 1.9: Best-fit results for the production signal strengths (a) and for the decay signal strengths (b) for the combination of ATLAS and CMS. Also shown for completeness are the results for each experiment. The error bars indicate the 1σ (thick lines) and 2σ (thin lines) intervals. Both figures are taken from Ref. [43].

based on the analysis of all the main production modes presented above and of the decay modes $H \to ZZ, WW, \gamma\gamma, \tau\tau, b\bar{b}, \mu\mu$; the main results are summarized in Figure 1.9, expressed as the ratio μ of the measured value over the SM prediction. These results are obtained under the assumption that the SM values for the Higgs boson BR are valid. These measurements provide the most precise and comprehensive experimental results on these quantities at the time of writing.

The combined signal strength was measured to be:

$$\mu = 1.09 \pm 0.11 \tag{1.38}$$

The aforementioned results were obtained considering as inputs the individual searches done by both collaborations. The relevant ATLAS inputs for the $t\bar{t}H$ production mode combine different searches exploit-

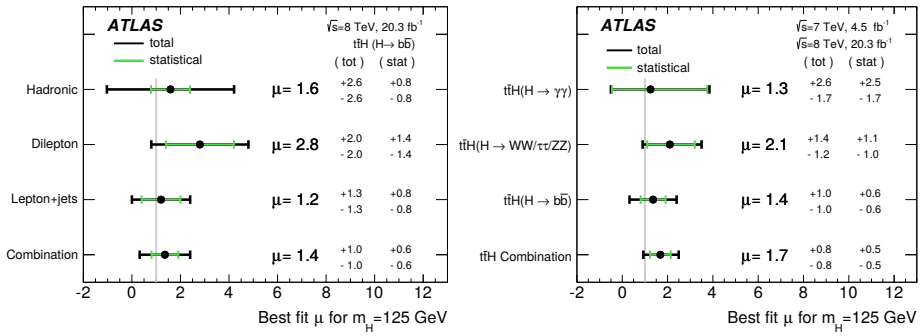
ing different final states ($H \to b\bar{b}$, $H \to (WW^*, \tau\tau, ZZ^*) \to$ leptons and $H \to \gamma\gamma$) and were performed at the centre of mass energy of 7 TeV and 8 TeV, corresponding to 4.5 fb^{-1} and 20.3 fb^{-1} respectively.

The result of the $t\bar{t}H(b\bar{b})$ combination for the signal strength was $\mu = 1.4 \pm 1.0$. The observed signal strengths for the individual $t\bar{t}H(b\bar{b})$ channels and for their combination are summarized in Figure 1.10.

The result for the best-fit value was:

$$\mu = 1.7 \pm 0.8 \tag{1.39}$$

The observed significance for the $t\bar{t}H(b\bar{b})$ combination is equal to 1.35 σ, whereas the observed (expected) significance of the combined $t\bar{t}H$ result is 2.33 σ (1.53 σ).



(a) $t\bar{t}H(b\bar{b})$ search **(b)** $t\bar{t}H$ combined search

Figure 1.10: Summary of the measurements of the signal strength μ for the $t\bar{t}H(b\bar{b})$ production for the individual $H \to b\bar{b}$ channels and for their combination (a) and summary of the measurements of the signal strength μ for the individual channels and for their combination (b). Both plots are done assuming m_H = 125 GeV. The total (tot) and statistical (stat) uncertainties of μ are shown. The SM expectation ($\mu = 1$) is shown as the grey line. Taken from Ref. [44].

The values reported by the ATLAS individual searches differ from those shown in Figure 1.9a due to small differences in, e.g., the mass of the Higgs boson used as input, small changes needed to perform the

combined measurement and mostly by the correlations and simultaneous contributions of the various analyses used as input in the fit.

1.3 Beyond the Standard Model

As it was already said, the Standard Model of Particle Physics is a well established model of great success, able to predict the results of particle physics experiments with extraordinary precision over a wide range of cross-sections, as can be seen in Figure 1.11, where several Standard Model processes cross sections are shown, spanning throughout 14 orders of magnitude.

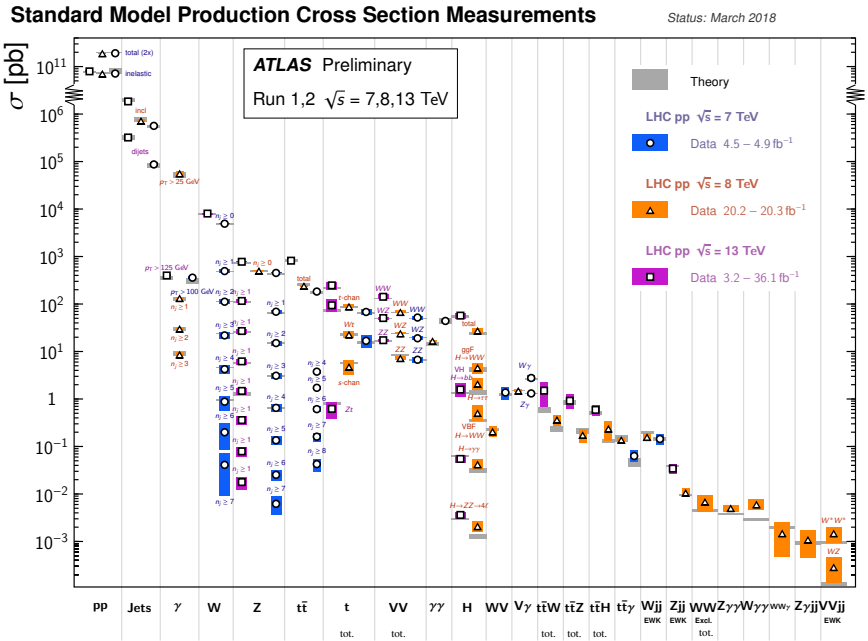


Figure 1.11: Summary of several Standard Model total and fiducial production cross section measurements, corrected for leptonic branching fractions, compared to the corresponding theoretical expectations. All theoretical expectations were calculated at NLO or higher. Not all measurements are statistically significant yet. Taken from Ref. [45].

Nevertheless, it has some intrinsic problems and does not provide satisfactory answers to several questions, among which there are:

hierarchy problem: the many orders of magnitude between the vector boson masses and the Planck scale[12].

origin of CP violation: even though the SM predicts the existence of CP violation, via the phase in the CKM matrix, it is not sufficient to explain the greater matter-antimatter asymmetry seen in the universe [46].

dark matter: the SM lacks a candidate to explain the large amount of dark matter seen in the universe.

origin of neutrino masses: in the SM the neutrinos are massless, but the observation of neutrino oscillations [9, 10] implies that these particles have a mass. A mass term for neutrinos could be included in the SM in an analogous way to the other leptons, but they could also be *Majorana* neutrinos, i.e. particles representing their own antiparticles, as they carry neither electric nor colour charge.

unification of forces: in the SM the strength of the fundamental forces varies depending on the energy of the process under consideration. Extrapolating their behavior up to the Planck scale does not present a meeting point for them, being a sign of no further unification, whereas some theories beyond the SM do show an unification.

gravity: it is not possible to include gravity in a coherent way in the Standard Model.

All these items suggest that the Standard Model is just a "low energy" limit of a more fundamental theory. Over the years, several possible extensions have been proposed. Among them SuperSymmetry (SUSY) has been considered for long the most elegant and natural extension of

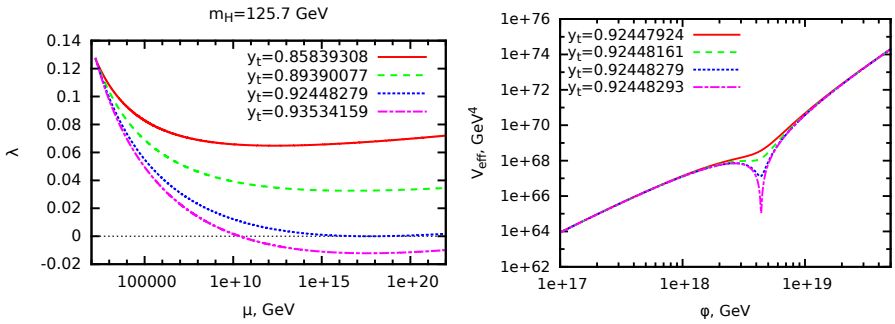
[12] The Planck scale is defined such as the scale at which the quantum and gravitational effects are of comparable size and it is approximately 10^{19}GeV.

the SM, as it is able to solve most of the problems exposed above. Sadly, as of today, no sign of new physics has been found at the LHC.

The absence of direct evidence for new physics poses more and more importance to indirect searches and tests of internal consistency of the SM, in the hope of finding significant deviations from the expected theoretical value of some quantity. In this regard, effects produced by new particles in loop processes are of extremely high importance, as well as investigations in the top and Higgs sector, since in most Beyond the Standard Model (BSM) theories new particles will couple strongly to the last generation of quarks or to the Higgs boson.

Nevertheless, even if the SM is valid up to a very high energy scale, studying the top Yukawa coupling and the Higgs sector is of extreme importance, as this coupling sheds light on the future of the universe, because it is not just one of the 19 free parameters of the theory. The renormalization group equation for the Higgs self-coupling shows a strong dependence on this coupling, which is the biggest in the SM, therefore changes at the percent level at low energy will translate into a huge difference at high energy, as can be seen in Figure 1.12.

Furthermore, the Higgs potential can cease to show the increasing monotonic behaviour it shows in the proximity of the minimum and develop a second minimum at higher energies, due to the running evolution of λ. The phase diagram in Figure 1.13 suggests that, given the current values of m_{t} and m_H, there may be a second, deeper minimum in the Higgs potential and, if that is the case, the universe could undergo a phase transition and move there by tunneling: the universe is metastable, meaning it is momentarily stable on cosmological time scales but inevitably headed towards a distant cataclysm in which life as we know it might be impossible. We may sit on the brink of an abyss [47].



(a) **(b)**

Figure 1.12: Renormalization group running of the Higgs coupling constant λ for the Higgs mass m_H = 125.7 GeV and several values of the top quark Yukawa y_t (left) and the effect of a very small change in y_t to the behaviour of the effective potential for the Higgs field to that with an extra minimum at large values of the Higgs field (right). Both plots have been taken from Ref. [48].

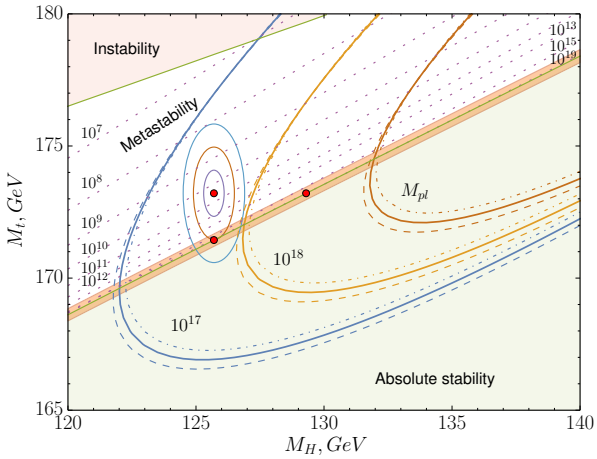


Figure 1.13: Phase diagram of vacuum stability (light-green shaded area), metastability, and instability (pink shaded area) in the (m_H,m_t) plane. The present world average of (m_H,m_t) (the upper left red bullet) is shown together with its 1σ (purple ellipse), 2σ (brown ellipse) and 3σ (blue ellipse) contours. Taken from Ref. [47].

The LHC and the ATLAS detector

2

At the end of the Second World War, European science was no longer world-class. A handful of visionary and pioneer scientists had the idea of creating a European nuclear physics laboratory, with the goal of not only sharing the costs for the facilities and establishing a world-class fundamental physics research organization in Europe, but also to unite European scientists.

On 17 May 1954, the first shovel of earth was dug on the Meyrin site in Switzerland, near the border with France in the Geneva area, and later that year, on 29 September 1954, following the ratification of all the founder states, CERN [1] came into being.

Started as one of Europe's first joint ventures, nowadays it has 22 member states and many more associated and observer states, making CERN one of the most successful international collaborations worldwide: over 600 institutes and universities around the world use CERN's facilities and more then 12000 visiting scientists from over 70 countries and with 105 different nationalities – half of the world's particle physicists – come to CERN for their research.

2.1 The Large Hadron Collider

The Large Hadron Collider (LHC) [49] is the most powerful particle accelerator ever built and the the latest addition to CERN's accelerator complex. It was fired up on 10 September 2008; at 10:28 am local time a beam of protons was successfully steered around the ring for the first time. It is built in a 26.7 km circumference tunnel located underground

[1] CERN is the acronym for the French "Conseil Européen pour la Recherche Nucléaire" or European Organization for Nuclear Research.

at a depth between 45 and 170 m on a plane inclined at 1.42% sloping towards the Geneva lake, which previously hosted the LEP accelerator.

Along the circumference of the LHC are placed four main detectors built around the four interaction points, ATLAS, CMS, LHCb and ALICE, and two smaller ones, TOTEM and LHCf:

ATLAS (A Toroidal LHC ApparatuS) [50, 51] is a multi-purpose detector designed to cover a wide range of physics searches. The main goals of the ATLAS experiment are precision measurements of the properties of the Standard Model particles, searches for new phenomena and new particles not included in the SM as well as the discovery of the Higgs boson.

CMS (Compact Muon Solenoid) [52] is the second multi-purpose detector built along the LHC ring. It has the same physics program as ATLAS and, even if its design features and some technical choices are opposite to the ones of ATLAS, it achieves comparable performance.

LHCb [53] is an experiment whose main physics program is the precise study of *b*-hadrons and CP violation in this sector of the SM.

ALICE (A Large Ion Collider Experiment) [54] is an experiment whose main purpose is the study of the phase transition in the quark-gluon plasma. This state of matter is achieved mostly by colliding lead-lead ion and proton-lead ion beams.

TOTEM (TOTal cross section, Elastic scattering and diffraction dissociation Measurement at the LHC) [55] is located at the two sides of the CMS experiment. Its physics program is dedicated to the precise measurement of the proton-proton interaction cross section, as well as to the in-depth study of the proton structure. The physics processes under study happen in the region very close to the particles beam (forward region).

LHCf (Large Hadron Collider forward) [56] is installed on both sides of the ATLAS experiment. Its aim is the study of the neutral particle production cross sections in the very forward region, in order

to provide insights for the development of atmospheric showers induced by very high energy cosmic rays.

The choice of a proton-proton collider was driven by the performance benchmarks needed to explore new frontiers in high energy physics. As a matter of fact, with proton collisions it is possible to explore a wide range of centre or mass energy, $\sqrt{s}$, while keeping constant the energy of the two colliding beams, due to the fact that the fundamental parton-parton interaction happens at a fractional energy of the proton-proton system ($\sqrt{\hat{s}} = \sqrt{x_1 x_2 s}$). Besides that, the higher mass of the proton compared to the mass of electrons and positrons reduces the energy loss due to emission of synchrotron radiation[2]. Lastly, it is not possible to produce anti-proton beams with an intensity sufficient to achieve the desired instantaneous luminosity.

During Run1, the LHC machine operated at a $\sqrt{s}$ of 7 TeV first and 8 TeV later, raised to $\sqrt{s}$ = 13 TeV in 2015 at the beginning of Run2. The increase in energy was possible because in the years between the two runs, the Long Shutdown 1 (LS1), all the electrical connections between the magnets were consolidated. The LHC machine can also collide lead ions at $\sqrt{s}$ = 5.02 TeV per nucleon. In order to reach such energies, the beams pass through the accelerator complex depicted in Figure 2.1.

For proton beams, the source is a simple bottle of hydrogen gas. Electrons are stripped from hydrogen atoms and the remaining protons are subsequently accelerated by Linac 2, the first accelerator in the chain, up to the energy of 50 MeV. The Proton Synchrotron Booster (PSB) accelerates the injected beam up to 1.4 GeV and then injects it in the Proton Synchrotron (PS), which pushes the beam to 25 GeV. Subsequently, the proton beam is sent to the Super Proton Synchrotron (SPS) to reach an energy of 450 GeV. It is eventually transferred and split into the two beam pipes of the LHC, where the two newly created beams circulate in opposite directions and their energy is ramped up to 6.5 TeV per beam.

Protons are injected in the LHC in the form of small *bunches*. Each

[2] This radiation is emitted in the form of photons when charged particles move along a curved trajectory. The energy loss per turn is proportional to $E^4/(m^4 R)$ which, for the LHC, translates into 6.7 keV per turn.

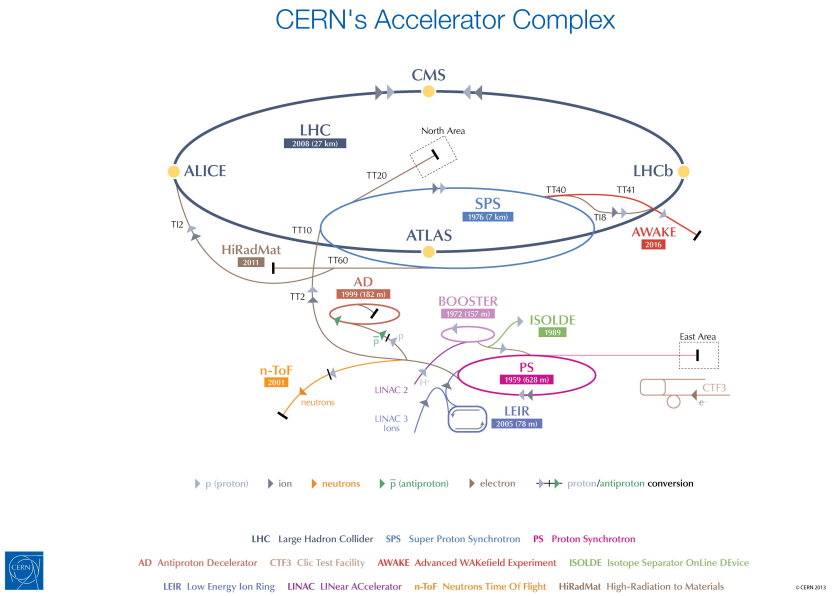


Figure 2.1: Schematic view of the CERN's accelerator complex [57].

bunch contains $\sim 10^{11}$ protons and every 25 ns a different bunch crossing happens. A system of 1232 dipole superconducting magnets with a bending field of 7.74 T is used to accelerate the beams in the LHC and "higher order" magnets (quadrupoles, ..., dodecapoles) are used to focus the beam. Just prior to the collision point, a system of quadrupoles is used to "squeeze" the particles closer together to increase the chances of collisions.

In total there are 9593 superconducting magnets all around the LHC machine, which are cooled down to a temperature of 1.9 K by superfluid helium, colder than outer space (2.7 K): the LHC is the largest cryogenic system in the world.

Table 2.1 contains the design and operating parameters of the LHC machine.

The instantaneous luminosity is one of the fundamental parameters of a collider. It is related to the number of events for a given process

Table 2.1: Overview of the LHC beam parameters comparing the design values with their operating values during the first two years of Run2.

	2015	2016	Design
Centre of mass energy [TeV]	13	13	14
Peak luminosity [10^{33} cm^{-2} s^{-1}]	5.0	13.8	10
Protons per bunch ($\times 10^{11}$)	1.2	1.1	1.15
Max bunches	2244	2220	2808
Mean interactions per crossing	13.7	24.9	23
Bunch crossing [ns]	50/25	25	25

produced per unit of time via the equation:

$$N_{\text{event}} = L\,\sigma_{\text{event}} \tag{2.1}$$

where σ_{event} is the cross section for the process under study and L is the machine luminosity, which depends only on the beam parameters. For a Gaussian beam distribution it can be written as:

$$L = \frac{N_b^2 n_b f_r \gamma_r}{4\pi\varepsilon_n \beta^*} F \tag{2.2}$$

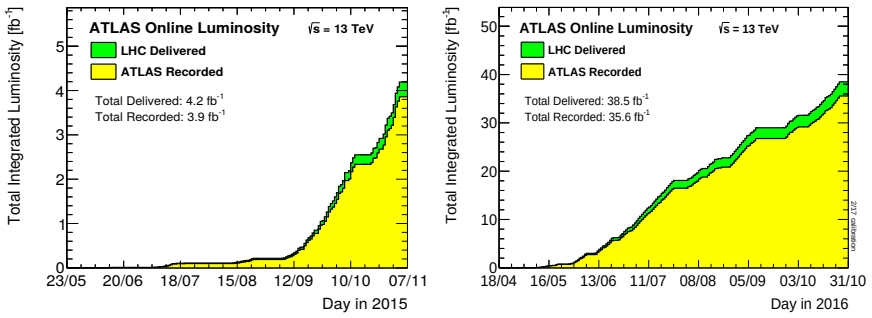
where N_b is the number of particles per bunch, n_b the number of bunches per beam, f_r the frequency of revolution, γ_r the relativistic gamma factor, ε_n the normalized transverse beam emittance, β^* the beta function at the collision point and F the geometric luminosity reduction factor due to the crossing angle of the beams at the interaction point.

The luminosity is not constant over a physics run, but decays due to the beam degradation. The main source of the luminosity decay during nominal LHC operation is caused by beam losses due to collisions, with a second contribution coming from particle losses due to a slow emittance blow-up, caused by the scattering of particles on residual gas and by beam-beam interactions. Taking into account all these effects, the luminosity lifetime, prior to the start of the machine, was estimated as:

$$\tau_L = 14.9\ \text{h} \tag{2.3}$$

a value met during actual LHC operations.

The total number of events produced is proportional to the luminosity integrated over time. The plots presented in Figures 2.2 and 2.3 are taken from the ATLAS luminosity public results for Run2 [58]. Figure 2.2 shows the total integrated luminosities delivered to and recorded by ATLAS during the 2015 and 2016 data taking campaign.



(a) **(b)**

Figure 2.2: Total integrated luminosities delivered to (green) and recorded by (yellow) ATLAS versus the day of the year in 2015 (left) and 2016 (right).

The delivered luminosity accounts for luminosity delivered from the start of stable beams until the LHC requests ATLAS to put the detector in a safe standby mode to allow for a beam dump or beam studies, whereas the recorded luminosity reflects the data acquisition inefficiencies.

In each bunch crossing, multiple pp interactions happen at the same time, an effect called *pile-up*. In Figure 2.3 the distribution of the mean number of interactions per bunch crossing is shown for all data delivered to ATLAS during stable beams, together with the individual profiles per year.

2.2 The ATLAS detector

ATLAS is a detector designed to have an excellent performance within a broad range of physics processes, which poses a series of stringent

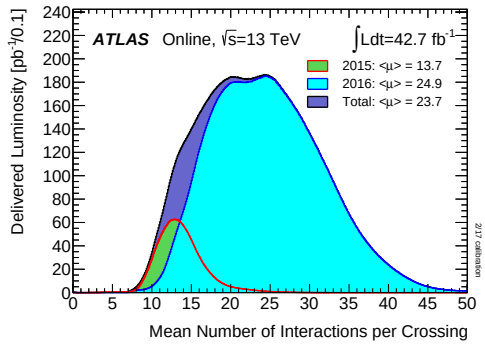


Figure 2.3: Distribution of the mean number of interaction per bunch crossing for the first two years of the Run2 collision data at $\sqrt{s}$ = 13 TeV, as well as the profile per each year of data taking.

requirements on its design, layout and capabilities. It means that the detector is able to efficiently reconstruct electrons, muons, tau leptons, photons and jets in a busy environment as a collision at LHC can be. All this translates into the need of:

- radiation-hard electronics and sensors with a fast time response in order to select events of interest and to suppress pile-up effects;
- maximal angular acceptance and hermeticity;
- excellent resolution on the momentum and other track parameters and high efficiency in reconstructing tracks left by charged particles, in order to be able to reconstruct the collision point, also known as *primary vertex* (PV), and extra displaced vertices;
- high granularity and resolution of both the electromagnetic and hadronic calorimeters;
- great muon identification and momentum resolution over a broad range of energies;
- a fast and efficient trigger system.

The momentum resolution of charged particles is proportional to:

$$\frac{\Delta p}{p} \sim \frac{p}{BL^2} \tag{2.4}$$

where B is the magnetic field inside the detector and L is the length of the arm used to measure the sagitta. High resolution can therefore be achieved with either an intense magnetic field over a small detector region or by having a big detector region immersed in a less intense magnetic field. CMS follows the first philosophy of construction, whereas ATLAS adopted the second one.

2.2.1 The magnet system

ATLAS features a hybrid system of four large superconducting magnets [59], which consists of:

- one solenoid in the barrel region, aligned with the beam axis, which provides a magnetic field of 2 T for the inner detector;
- one barrel toroid, which surrounds both calorimeters and both end-cap toroids and produces a field with an average value of 0.5 T and a peak field value of 3.9 T;
- two end-cap toroids, which produce a magnetic field with an average value of 1 T, required to optimize the bending power in the end-cap regions of the muon spectrometer system, with a peak field value of 4.1 T.

The layout of the barrel solenoid was carefully optimized to keep the material thickness in front of the calorimeters as low as possible, resulting in 0.66 radiation lengths at normal incidence, with the inner and outer diameters of the solenoid that are of 2.46 m and 2.56 m and a total length of 5.8 m. The size of the barrel toroid is 25.3 m in length, with inner and outer diameters of 9.4 m and 20.1 m respectively, while the end-cap toroids have a inner and outer diameters of 1.65 m and 10.7 m and a length of 5.0 m. A schematic view of the three parts of ATLAS magnet system is depicted in Figure 2.4.

These characteristics of the magnet system define the geometry and the structure of the ATLAS detector, as shown in Figure 2.5. Like most detectors, it has an approximately cylindrical and a forward-backward symmetry with all its sub-detectors positioned in concentric layers. The closest sub-detector to the beam pipe is designed to reconstruct the tracks of charged particles, followed by the electromagnetic and hadronic calorimeters, whose task is to measure the energy of photon, electrons and hadrons. The last sub-detector is designed to identify muons and measure their energy.

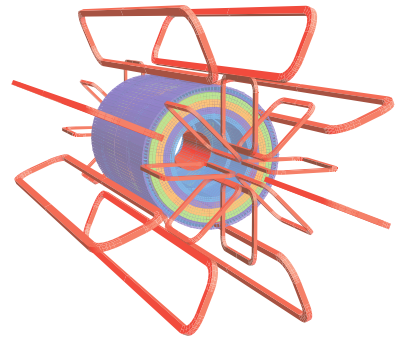


Figure 2.4: Schematic view of the ATLAS magnetic system.

2.2.2 Coordinate system

ATLAS uses a right-hand coordinate system with its origin at the nominal interaction point (IP) in the centre of the detector and the z-axis along the beam direction, with positive z in the anti-clockwise direction. The x-axis points to the centre of the LHC ring and the y-axis points upward.

A cylindrical coordinate system is employed, defined by:

- $R = \sqrt{x^2 + y^2}$, the distance from the beam line in the x-y plane, also called transverse plane;
- $\phi \in [-\pi; \pi]$, the azimuthal angle measured in the transverse plane;
- $\theta \in [0; \pi]$, the polar angle measured with respect to the beam axis.

It is common to re-express the polar angle as the *pseudorapidity* η:

$$\eta = -\ln \tan \frac{\theta}{2} \tag{2.5}$$

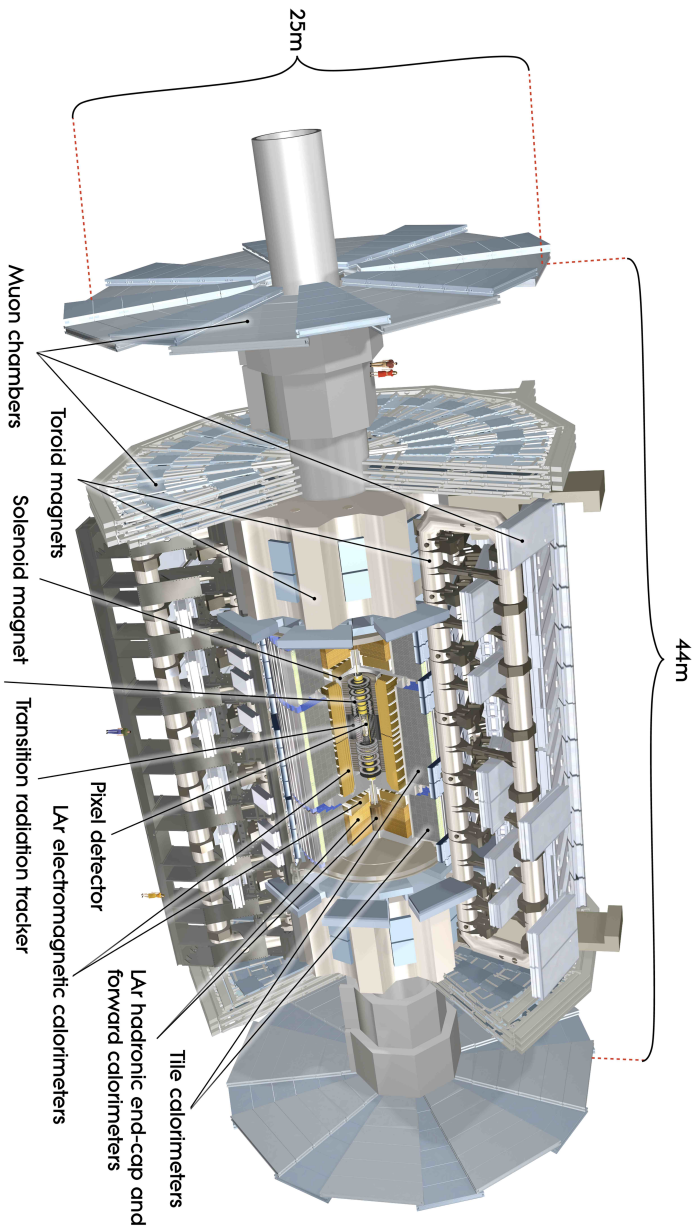


Figure 2.5: Schematic view of the ATLAS detector and its different sub-detectors [60].

which is the limit for massless particles of the rapidity $y = \ln \frac{E+p_z}{E-p_z}$. This is done because (pseudo)rapidity differences are a Lorentz invariant quantity for boosts along the beam direction.

The angular distance between two objects is commonly expressed as:

$$\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2} \tag{2.6}$$

2.2.3 The inner detector

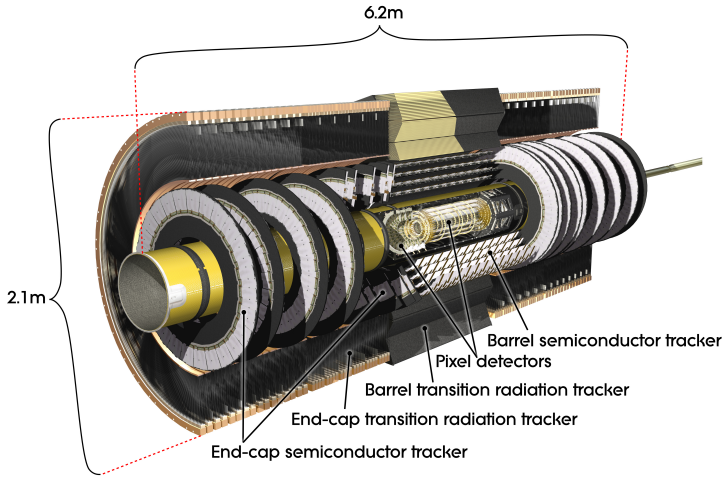


Figure 2.6: Computer generated image of the ATLAS inner detector [61]. Moving away from the beam pipe the Pixel Detector is found, followed by the SCT and TRT sub-systems.

The Inner Detector (ID) [62] provides charged particle tracking with high efficiency over the range of $p_{\mathrm{T}} > 0.4$ GeV and $|\eta| < 2.5$.

High spatial resolution is of vital importance in order to reconstruct tracks in the busy environment of a collision at the LHC, as well as for the ability to reconstruct displaced vertices. Therefore, in order to resolve different tracks in a very large track-density environment, a fine granularity around the vertex region is needed. The ATLAS Pixel Detector and the Semiconductor Tracking (SCT) offer these features.

The third sub-detector encountered moving away from the beam pipe is the Transition Radiation Tracker (TRT), which provides a large number of tracking points. The combination of the these elements gives very robust pattern recognition and high precision in both ϕ and z.

In the barrel region ($|\eta| < 1.0$), the elements are arranged on concentric cylinders around the beam axis, while in the end-cap regions they are placed on disks perpendicular to the beam axis. The layout of the ID can be seen in Figure 2.6, while in Figure 2.7 it's possible to see also the distance of the various elements of the barrel from the beam line.

Overall, the resolution of the tracking system is:

$$\frac{\sigma_{p_T}}{p_T} = 0.05\% p_T \oplus 1\% \tag{2.7}$$

where $\oplus$ denotes the sum in quadrature of the two terms.

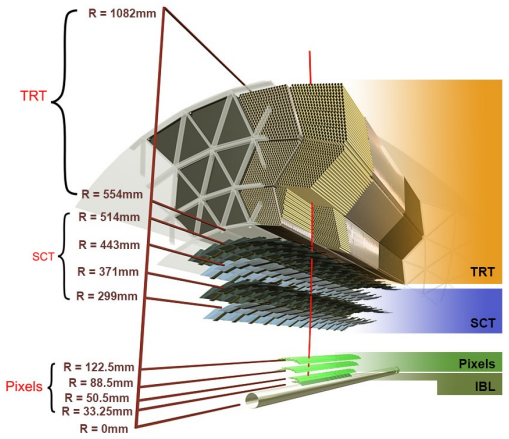


Figure 2.7: Sketch of the ATLAS inner detector showing all its components, including the new IBL, taken from Ref. [63]. The distances with respect to the interaction point are shown as well.

The pixel detector

The innermost part of the ID is composed of 3+1 silicon pixel layers organized in a cylindrical manner in the barrel and three disks in the end-caps. The Insertable B-Layer (IBL) is an addition to the already existing three layers before the start of Run2 [64], as a measure to cope with the increased pile-up conditions and to improve the vertexing and *b*-tagging performance. In fact, the IBL improves tracking by providing an additional measurement point and mitigates the possible loss of hits in the other three layers, which can happen with the high integrated luminosity and after radiation damage. In order to be able to include the new IBL, the beam pipe had to be redesigned and made smaller by 4 mm that, added to the space left empty in the previous design, left 12.5 mm to insert the IBL.

The IBL is installed at an mean radius of 33.2 mm around the new beam pipe and it is equipped with pixel sensors with size (ϕ, z) of (50, 250) μm, providing a resolution of 10.0 μm and 66.5 μm for $R-\phi$ ($|\eta| < 2.0$) and z (for $|\eta| < 1$) directions respectively [65].

The old pixel detector layers are located at radius of 50.5 (B-Layer), 88.5 (Layer 1) and 122.5 mm (Layer 2) away from the beam axis. Each piece of each layer is slightly tilted and overlaps with the adjacent neighbours both in the longitudinal and radial direction, in order to ensure perfect hermeticity, thus each track is expected to hit all the layers. The typical pixel size is 50 $(R-\phi)$ $\times$400 (z) μm and the sensor has a thickness of 250 μm each. In the barrel, the intrinsic resolution is 10 μm $(R-\phi)$ and 115 μm (z), whereas in the disks this is 10 μm $(R-\phi)$ and 115 μm (R).

In total, the Pixel Detector has approximately 86 million channels.

Semiconductor Tracking

The semiconductor tracking is a silicon strip detector composed of four cylindrical layers in the barrel and nine disks in the end-caps. The SCT modules are arranged such that each charged particle crosses eight strip layers to give rise to four space points.

They consist of two 6.4 cm long sensors with a strip pitch of 80 μm, for an intrinsic resolution of each module, in the barrel, of 17 μm in

$R-\phi$ and 580 μm in z and a resolution of 17 μm in $R-\phi$ and 580 μm in R for the end-cap disks.

The total number of readout channels in the SCT is approximately 6.3 million

Transition Radiation Tracker

The outermost part of the ID is a straw tube detector, which covers the pseudorapidity range $|\eta| < 2.0$. Each straw has a 2 mm radius and is filled with a gas mixture (3% O_2 , 27% CO_2 and 70% Xe).

When a relativistic charged particle travels through an inhomogeneous material, it emits transition radiation proportional to the Lorentz factor $\gamma = E/m$. This radiation subsequently ionizes the gas mixture creating a current; based on the amount of current produced it is therefore possible to distinguish particles with the same energy but different masses, such as pions and electrons. The maximum drift time of the induced current is of 40 ns. The space between the tubes is filled with polyethylene, which serves as the material for the transition radiation.

The straw tubes are 144 cm long and parallel to the beam axis in the barrel, with the wire electrically divided in two halves at $\eta = 0$, while they are 37 cm long and have a radial configuration in the end-caps, providing information only in the $R-\phi$ plane, for which it has an intrinsic accuracy of 130 μm per straw.

The lower resolution is mitigated by the large number of hits per track (typically 36 per track) and the longer measured track length. The total number of TRT readout channels is approximately 351000.

2.2.4 The calorimeter system

Calorimeters are used to measure the energy of a particle. The ATLAS calorimetric system is built with two types of them, each designed to target the detection of different particles, namely electromagnetic (EM) particles (electrons and photons) and hadrons.

Both the electromagnetic (ECAL) and the hadronic (HCAL) calorimeters are *sampling* calorimeters, built with alternating layers of absorber and active material. When a particle travels through the absorber, it heavily interacts with it and creates a shower, detected by the active

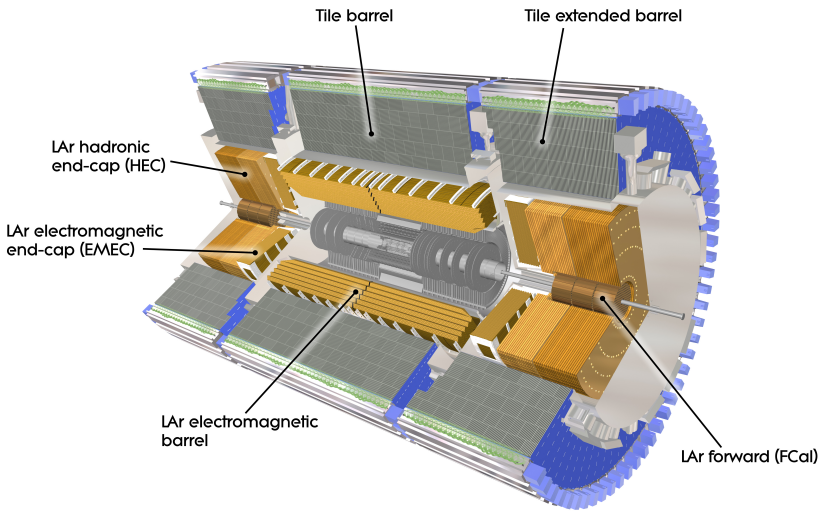


Figure 2.8: Cut-away view of the ATLAS calorimeter system: the Liquid Argon (LAr), the Tile and the forward calorimeters [66].

material; this process continues at each layer, until all the energy of the particle is fully deposited in the calorimeter, hence, the depth of a calorimeter is an important design parameter.

These calorimeters cover the range up to $|\eta| < 4.9$, using different techniques that are suited for the different requirements due to the radiation environment over the large η range. In the η range matched by the ID, the fine granularity of the ECAL allows for a precision measurement of electrons and photons, whereas the coarser granularity of the rest of the calorimeter is sufficient to meet the requirements for a precise jet and $E_{\mathrm{T}}^{\mathrm{miss}}$ reconstruction.

Liquid Argon (LAr) is chosen as the active material for the EM calorimeter and the hadronic calorimeters in the end-cap and forward regions for its intrinsic linear behaviour, its stability of response over time and its intrinsic radiation-hardness; on the other hand, a different technology for the barrel hadronic calorimeter was chosen as it can provide maximum radial depth for the least cost.

A view of the sampling calorimeters is presented in Figure 2.8.

Electromagnetic calorimeter

The ECAL [67] barrel region extends up to $|\eta| < 1.475$ and the two end-caps have a range of $1.375 < |\eta| < 3.2$. The barrel calorimeter consists of two identical half-barrels, separated by a small gap of 4 mm at $z = 0$.

In the region of $|\eta| < 1.8$, a *pre-sampler* detector is used to correct for the energy lost by electrons and photons in the solenoid, before reaching the calorimeter.

The absorber is made out of lead and it has a characteristic accordion shape[3], depicted in Figure 2.9a.

It is segmented into three longitudinal layers. The first layer (called *strip layer*) is finely segmented in η (0.0031) for a precise determination of the shower properties, in particular for photon and electron identification, and distinguish photon pairs coming from a pion decay from promptly produced photons. The second layer is the one with the larger thickness (16 radiation lengths, X_0) and is where the majority of the energy is deposited; with granularity of 0.025×0.025 in terms of $\Delta\eta \times \Delta\phi$ cell size for the barrel and 0.050×0.025 for the end-caps. The third layer collects only the tail of the electromagnetic shower and is therefore not finely segmented in η. The total thickness of the EM calorimeter is of a minimum of 22 X_0 in the barrel and 24 X_0 in the end-caps, increasing with η.

The designed energy resolution, for the barrel ECAL after electronic noise is subtracted, is given by [51]:

$$\frac{\sigma_E}{E} = \frac{10.1\%}{\sqrt{E[\text{GeV}]}} \oplus 0.17\% \tag{2.8}$$

Hadronic calorimeter

ATLAS hadron calorimetry is also based on the sampling technique. In the central region the *Tile* calorimeter [68] is placed just outside ECAL and it covers the regions of $|\eta| < 1.0$ in the barrel and $0.8 < |\eta| < 1.7$

[3] The accordion geometry has been chosen as it provides complete ϕ symmetry without any cracks. By placing the accordion waves with different angles and distances among themselves, depending on the position of the element, it is possible to achieve a very uniform performance in terms of linearity and resolution as a function of ϕ.

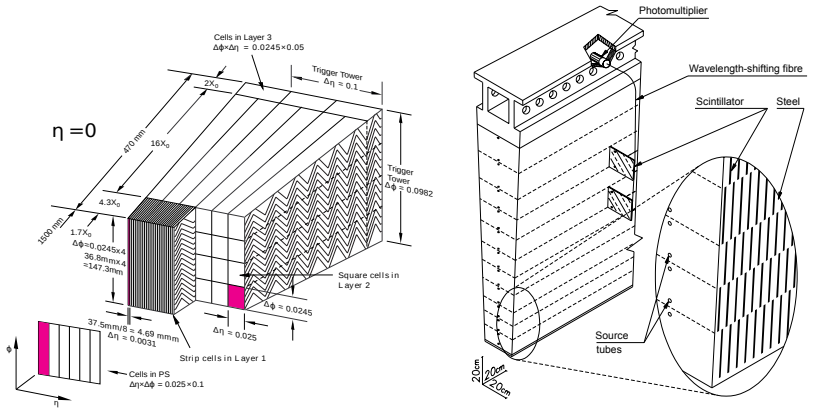


Figure 2.9: Sketch of (a) an ECAL barrel module, where the different longitudinal layers are depicted with the characteristic accordion shape and (b) a Tile calorimeter module, also showing the layers of steel absorber and active scintillator material, as well as the optical read-out.

in the extended barrel. It uses steel as absorber and scintillating tiles as active material, arranged as in the structure shown in Figure 2.9b. It is segmented in depth in three layers of 1.5, 4.1 and 1.8 interaction lengths (λ) thick for the barrel and 1.5, 2.6, and 3.3 λ for the extended barrel. The total thickness is sufficient to reduce punch-through, i.e. the leakage of shower particles into the muon spectrometer, well below the irreducible level of prompt or decay muons. In the barrel region, the granularity is equal to 0.1×0.1 in $\Delta\eta \times \Delta\phi$ in the first two layers and 0.1×0.2 in $\Delta\eta \times \Delta\phi$ in the third.

Two independent wheels per end-cap constitute the Hadronic End-cap Calorimeter (HEC), placed directly behind the ECAL end-caps. Each wheel is divided into two segments in depth. It covers the range $1.5 < |\eta| < 3.2$ and uses copper as absorber and LAr as active material. The granularity is equal to $\Delta\eta \times \Delta\phi = 0.1 \times 0.1$ in the region $|\eta| < 2.5$ and $\Delta\eta \times \Delta\phi = 0.2 \times 0.2$ for higher $|\eta|$.

The LAr Forward Calorimeter (FCal) provides full coverage over the range $3.1 < |\eta| < 4.9$. As it is located at high η, FCal is exposed to

high particle fluxes: the use of LAr allows for constant change of the active material which results in a more radiation-hard detector. It is approximately 10 λ in depth and consists of three modules per end-cap: the first is made of copper and is optimized for EM measurements, while the other two are made of tungsten, for hadronic interactions.

The energy resolution for Tile and HEC is:

$$\frac{\sigma_E}{E} = \frac{50\%}{\sqrt{E}} \oplus 3\% \tag{2.9}$$

whereas, for FCal it is:

$$\frac{\sigma_E}{E} = \frac{100\%}{\sqrt{E}} \oplus 10\% \tag{2.10}$$

2.2.5 The muon spectrometers

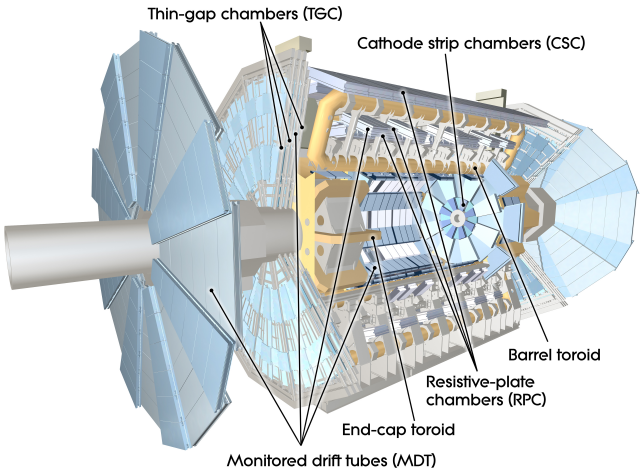


Figure 2.10: Schematic view of the ATLAS muon spectrometer [69].

The Muon Spectrometer (MS) [70] is the outermost sub-detector of ATLAS, designed to detect charged particles exiting the calorimeters and measure their momentum up to pseudorapidity $|\eta| < 2.7$ with a resolution better than 3% over a wide p_{T} range and up to 10% for a

particle of 1 TeV. It consists of one barrel ($|\eta| < 1.05$) and two end-caps ($1.05 < |\eta| < 2.7$). The material between the interaction point and the MS ranges approximately from 100 to 190 radiation lengths depending on η, and consists mostly of the calorimeters, which translates into a low limit of $p_{\mathrm{T}} \sim 3$ GeV for momenta measured by the MS alone.

As can be seen in Figure 2.10, the titanic size of ATLAS is defined by the MS. It has four detection systems, each one exploiting different technologies in order to have fast detectors for triggering and precision chambers for track reconstruction.

In the barrel region, tracks are measured in chambers arranged in three layers organized in a cylindrical fashion around the beam axis at $R \approx 5$ m, 7.5 m and 10 m. In the end-cap regions, the chambers form large wheels perpendicular to the z axis and located at distances of $|z| \approx 7.4$ m, 10.8 m, 14 m, and 21.5 m, as shown in Figure 2.11.

The four detection systems are:

MDT (Monitored Drift Tube) chambers, that, as the name suggests, consist of three to eight layers of drift tubes. They provide precise momentum measurements in the full $|\eta|$ range covered by the MS, with a precision of about 35 μm per chamber.

CSC (Cathode Strip Chambers) are multi-wire proportional chambers with cathodes segmented into strips in orthogonal directions. The high radiation resistance, high rate capability and time resolution are ideal for their use in the innermost plane ($2 < |\eta| < 2.7$). The spatial resolution is 40 μm in the bending plane and about 5 mm in the transverse plane.

RPC (Resistive Plate Chambers) are used in the barrel for triggering, as their time resolution of a few ns is below the bunch crossing time of 25 ns, but the drawback is their spatial resolution of only 10 mm. They are gaseous parallel electrode-plate detectors.

TGC (Thin Gap Chambers) are used in the end-cap wheels (but up to $|\eta| < 2.4$ for triggering). They are multi-wire proportional chambers with the characteristic that the wire-to-cathode distance is smaller than the wire-to-wire distance, leading to a very good time resolution similar to the RPC and a spatial resolution of a few mm.

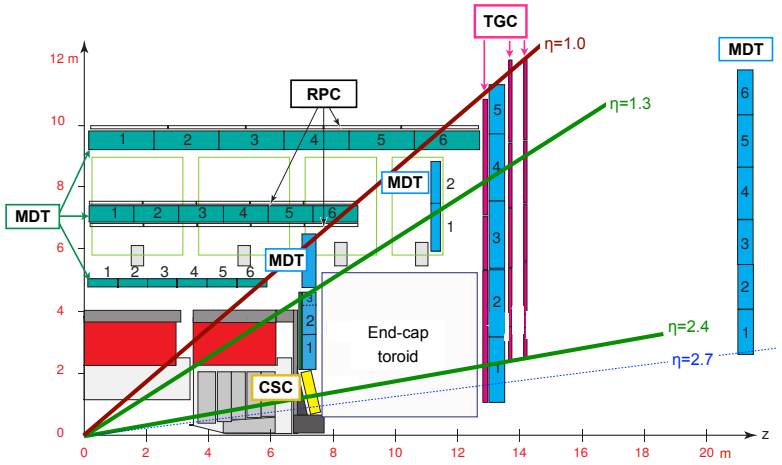


Figure 2.11: Cross-section of ATLAS muon system in the $R-z$ projection at $\phi = \pi/2$ (bending plane). Figure taken from Ref. [71].

2.2.6 The trigger system

One of the biggest challenges for the LHC experiments is the online selection of the events of interest, as it is impossible to record all the events produced. For this reason the trigger system plays a crucial role because it is responsible for deciding whether or not to keep an event for permanent storage.

During the Long Shutdown 1, the ATLAS trigger system [72, 73] was upgraded in order to cope with the foreseen increase in luminosity and number of interactions per bunch-crossing. Figure 2.12 depicts the layout of the ATLAS trigger and data acquisition system (TDAQ) in Run2. It is composed of two levels, the hardware based first level trigger (L1) and the software based high-level trigger (HLT).

The L1 trigger has two sub-triggers, the L1Calo and L1Muon, that determine Regions-of-Interest (RoIs) in the detector by taking as input coarse granularity calorimeter and muon detector information. The L1 trigger decision is taken by the Central Trigger Processor (CTP) and the event rate is reduced from the bunch crossing rate of 40 MHz down to 100 kHz.

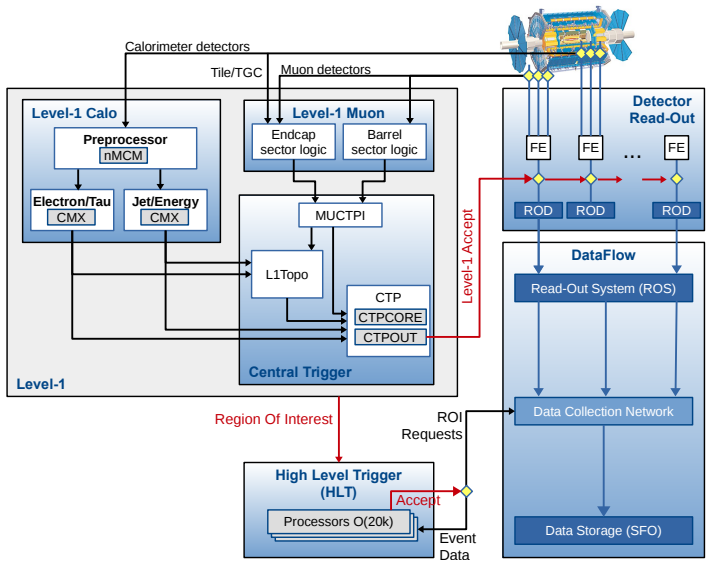


Figure 2.12: Schematic layout of the upgraded ATLAS trigger and data acquisition system in Run2. Adapted from Ref. [73].

In case the event is accepted, the CTP distributes the L1 accept signal and LHC timing signals to the sub-detector readout systems via the Timing, Trigger and Control network. The event information is stored in the front-end (FE) memory buffers located on the detector, while waiting for the L1 decision. After the L1 accept, the events are transferred to the Read-Out Drivers (ROD) located outside the detector, buffered in the Read-Out System (ROS) and processed by the HLT.

The HLT performs a fast reconstruction guided by the RoIs created by the L1, with software very close (or identical) to the offline algorithms. It uses finer granularity for the calorimeter information, precise measurements from the MS and tracking information from the ID, typically using information within an RoI identified by L1. Most of the trigger reconstruction algorithms were re-optimized during the shutdown to minimize differences between the HLT and the offline analysis selections, which in some cases reduced inefficiencies by more than a factor two, e.g., in the case of hadronic tau triggers.

The HLT reduces the rate from the Level 1 output rate to approximately 1 kHz. After the events are accepted by the HLT, they are trans-

ferred to local storage at the experimental site and exported to the Tier-0 facility at CERN's computing centre for offline reconstruction.

In the L1 Central Trigger, a new topological trigger (L1Topo) allows to apply topological selections at the L1 stage, combining kinematic information of the trigger objects received from the L1Calo or L1Muon systems, such as angular separation, invariant mass requirements or global event quantities like $E_{\mathrm{T}}^{\mathrm{miss}}$ [72]. The use of L1Topo significantly suppresses backgrounds for many trigger selections, in many cases by more than a factor of two, allowing to preserve the sensitivity for channels like $B_s \to \mu\mu$ at the current level in spite of the increasing luminosity. This system was fully installed and commissioned during 2016, but it was not used to collect the data used in this thesis.

2.2.7 Luminosity measurement

An accurate measurement of the delivered luminosity is a fundamental ingredient for many ATLAS analyses, such as cross-section measurements, for which the uncertainty in the delivered luminosity is often one of the major systematic uncertainties.

The total instantaneous luminosity is given by the following expression:

$$L = \sum_{b=1}^{n_b} L_b = n_b \langle L_b \rangle = n_b \frac{\mu_{\mathrm{vis}} f_r}{\sigma_{\mathrm{vis}}} \tag{2.11}$$

where the sum runs over the n_b bunch pairs colliding at the interaction point, $\langle L_b \rangle$ is the mean bunch luminosity, f_r is the bunch revolution frequency, μ_{vis} is the visible interaction rate per bunch-crossing and σ_{vis} is the visible inelastic cross-section, i.e. the total inelastic cross-section multiplied by the efficiency and the acceptance.

ATLAS monitors the delivered luminosity by measuring μ_{vis} and σ_{vis} with a variety of detectors and algorithms. These multiple detectors and algorithms are characterized by significantly different acceptances, response to pile-up, sensitivity to instrumental effects and to beam-induced backgrounds.

Each detector and algorithm is calibrated by determining its σ_{vis} using the so-called van der Meer scans [74], which are special low-intensity

LHC runs whose beam separation in the transverse planes is scanned, i.e. varied, in order to determine the overlap profile of the beams.

The primary detectors for luminosity measurement are:

BCM (Beam Conditions Monitor) consists of four 8×8 mm^2 diamond sensors arranged around the beam-pipe in a cross pattern manner at $z = \pm 1.84$ m on each side of the ATLAS IP, at $|\eta| = 4.2$. They are used both to detect beam instabilities and losses to provoke a beam dump, in order to avoid damage to the detector, and to get luminosity information for each bunch crossing, given the fast electronics (sub-nanosecond time resolution).

LUCID (LUminosity measurement using a Cherenkov Integrating Detector) consists of sixteen aluminium tubes filled with gas surrounding the beam-pipe on each side of the IP at a distance of 17 m, covering the pseudorapidity range $5.6 < |\eta| < 6.0$. If one of the photomultipliers produces a signal over a pre-set threshold, that tube records a hit for that bunch crossing.

Furthermore, ATLAS has additional ways of measuring the luminosity using its sub-detectors. In particular, the luminosity is assumed to be proportional to the number of reconstructed tracks found in the ID or the particle flux from pp collisions, which is monitored by the total ionization current flowing through a chosen set of LAr cells or by the current in Tile photomultiplier tubes. The drawbacks of these methods is that a bunch-by-bunch luminosity measurement is not possible.

A more detailed description of the methods can be found in Ref. [75].

2.2.8 Monte Carlo simulation

The generation of simulated events using Monte Carlo (MC) techniques is used in ATLAS for a wide variety of purposes, such as the development of new algorithms or performance studies, as well as for more general physics studies.

Monte Carlo programs generate events for a specific physics process in four steps: event generation, detector simulation, digitization and event reconstruction. First, the pp collision events are generated, including the hard scattering part, the showering of partons and the sub-

sequent hadronization evolution of the outgoing particles, including the underlying event as well. Such events are called *truth-events*.

The next step is the simulation of the interactions of these truth particles with the detector, how they shower into secondary particles and the energy deposit in each of the sensitive materials. The simulated energy deposit is then transformed into the detector response. The full ATLAS detector simulation [76] is based on the GEANT4 program [77]. A faster, less CPU-expensive simulation called *Atlfast-II* or AFII [78] exists as well. The reduction in computing time comes at the expenses of a less precise detector response, as the calorimeter simulation is replaced by a detailed parametrization of the shower shapes. The simulated events are reconstructed using the same reconstruction software used for real data. Generated MC events that underwent this chain are called *reco-events*.

There are two main types of MC generators: matrix element and general purpose generators. The former compute only the hard scattering part at a certain level in the perturbation theory (typically leading order or next to leading order) and can include a specified number of initial and final state radiation partons, but have to rely on the general purpose generators for the parton showering process, given that this type of generator can generate the entire proton-proton collision.

SHERPA [79], Herwig [80, 81] and PYTHIA [82] calculate both the matrix element and the parton shower, but take into account only $2 \to 1$ and $2 \to 2$ processes. On the contrary, MADGRAPH5_aMC@NLO [83] and POWHEG [84–87] can compute $2 \to n$ processes at next-to-leading order accuracy, but need to be interfaced with other generators to perform the showering.

In case of a LO matrix element generator, the process of interfacing it with a parton shower is a straightforward operation, because all the additional partons present in the final state will come directly from the showering algorithm. This is not the case for a NLO matrix element generators, as care has to be taken in order to avoid double counting of additional radiation that can be produced either by the matrix element itself or by the parton shower generator.

Given the complexity of (non-)perturbative QCD processes, MC generators used for the description of the parton shower typically employ

either some approximations for the high-multiplicity perturbative QCD calculations or phenomenological attempts to model non-perturbative effects that are not understood from first principles. They are based on empirical models, therefore various parameters can be adjusted in order to have a better description of real data. This is the so-called tuning of a generator, which in turn is simply the optimization process [88, 89].

Objects definition 3

The subsequent step after the recording of the event, whether it is a real data event or a simulated one, is the offline reconstruction. Information of the sub-detectors in the form of hits and energy deposits is grouped together in order to reconstruct and identify physics objects, such as leptons or jets.

It is possible to identify the various physics objects because each type of particle leaves a different trace in the various sub-detectors, as depicted in Figure 3.1. Charged particles leave tracks in the inner detector and then are stopped by one of the calorimeters to measure their energy. Muons are able to traverse the whole apparatus and leave a track in the outermost muon spectrometer, whilst neutrinos escape the detector without leaving any trace of their passage, hence their production is inferred from a momentum imbalance in the transverse plane.

3.1 Tracks and vertices

Hits in the inner detector are combined into reconstructed tracks using a sequence of algorithms, referred to as *New Tracking* (NEWT) [90, 91]. It is based on two main approaches:

inside-out track finding starts with the search for triplets of points in the silicon detectors (pixel or SCT), called seeds. Tracks using these seeds are then extended to the outer layers of the ID, with the additional points chosen by a combinatorial Kalman filter technique within the silicon detectors [92]. Tracks are then subject to an ambiguity solving algorithm that aims at eliminating candidates due to combinatorics, *fakes*, and the surviving candidates are then extended into the TRT.

outside-in is subsequent and complementary to the previous approach, as not all tracks can be found with it, because, e.g., tracks from photon conversions or secondary decay vertices stem from a point inside the ID and may not have enough or any silicon hits to form a seed. In this case, a reverse sequence is used, starting from a global pattern recognition in the TRT and tracing back the segments in the ID.

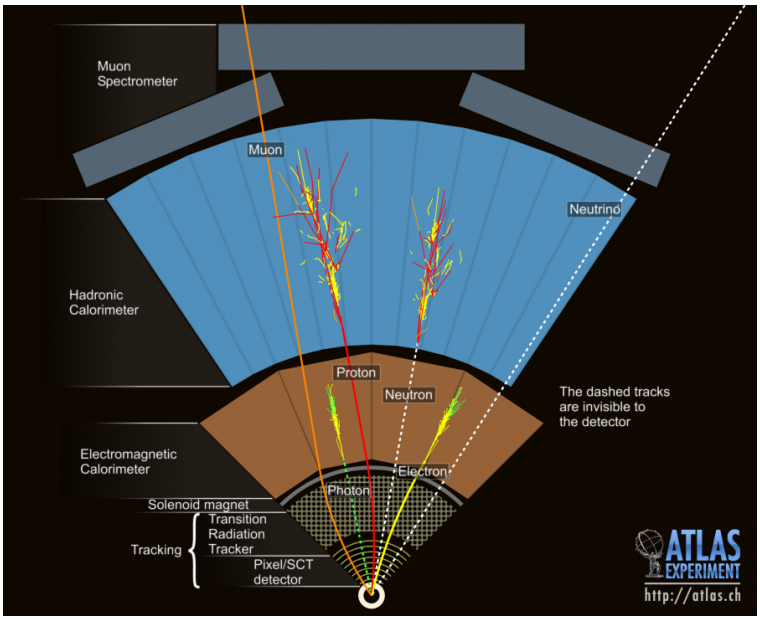


Figure 3.1: Overview of passage of particles through the ATLAS detector. Each physics object leave a different trace in the various sub-detectors, allowing for the particle identification [93].

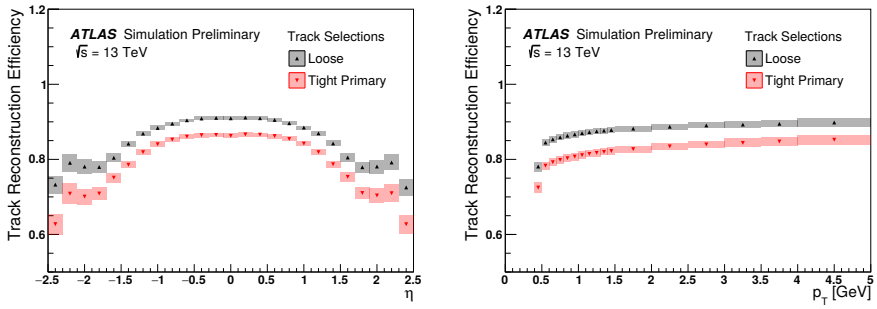
A reconstructed track is fully specified by five parameters, used to describe the helix: the transverse impact parameter, d_0, which is defined as the shortest distance between a track and the primary vertex (PV) in the plane transverse to the beam line; the longitudinal impact parameter, z_0, defined as the track z coordinate at its point of closest approach to the beam line with respect to the PV z coordinate; the angles ϕ and θ, which describe the direction of the track; and the ratio q/p, which

combines the momentum and charge of the track.

Quality cuts are later applied to the track parameters and to the number of hits, holes[1] and shared hits in sub-detectors in order to improve the object definition and the selection performance, improve the resolution and reduce the fake rates.

All reconstructed tracks must have at least a p_{T} larger than 400 MeV and $|\eta| \leq 2.5$. Two different classifications exist for tracks: *Loose* tracks require at least seven hits, with a maximum of two holes in the silicon detectors (Pixel and SCT) but at most one hole in the Pixel detector, and at most one shared module[2]; and *Tight Primary* tracks, defined by requiring, in addition to the *Loose* track criteria, at least nine (eleven) silicon hits in the region $|\eta| \leq 1.65$ ($|\eta| \geq 1.65$), one hit in the IBL or the B-Layer and no pixel holes.

The tracking efficiency is shown in Figure 3.2 as a function of p_{T} and η for both selections [94]. The *Loose* track selection is the default one.



(a) **(b)**

Figure 3.2: Track reconstruction efficiency, evaluated by using minimum bias simulated events, as a function of truth η (a) and p_{T} (b) for *Loose* and *Tight Primary* track selections. The bands indicate the total systematic uncertainty. Both plots are taken from Ref. [94].

A vertex is defined as the space point obtained by the intersection of several tracks, stemming from it. Due to the large number of protons

[1] A hole is essentially a missing hit, i.e. a hit in the detector was expected but none was found.

[2] A pixel module is considered to be shared if it has one or more shared hits, while a shared module in the SCT has at least two shared hits.

per bunch, several interaction vertices are reconstructed per event. The *primary vertex*, representing the interaction point of the hardest proton-proton collision, is generally assumed to be the one with the highest $\sum p_{\mathrm{T}}^2$. The kinematics of the physics objects are then recomputed considering this vertex as a new reference point. Figure 3.3 shows the primary vertex reconstruction efficiency as a function of the number of tracks associated with the vertex; the efficiency reaches 100% for vertices with at least four tracks.

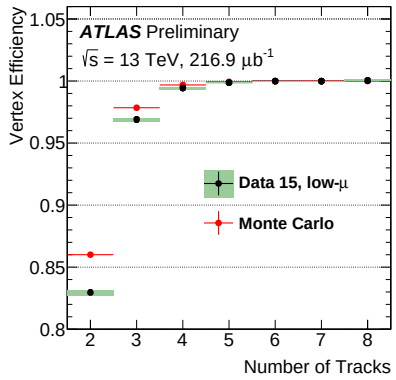


Figure 3.3: Vertex reconstruction efficiency as a function of the number of tracks, based on a small subset of a low-μ dataset [95].

3.2 Leptons

In this section the reconstruction and identification of electrons and muons will be discussed. On the other hand, τ-lepton reconstruction and identification will not be discussed as they are not explicitly used in the analyses discussed in this thesis. Instead, their leptonic decays are treated as electron or muons, whereas the hadronic decay topology will be reconstructed as a jet with specific properties [96].

3.2.1 Electrons

The reconstruction of electrons makes use of information from both the ID and ECAL sub-detectors, hence electron candidates are reconstructed only in the central region of the detector, up to $|\eta| < 2.47$ [97].

Electron reconstruction starts by looking for energy deposits (clusters) in the EM calorimeter with size 3×5 in units of $\Delta\eta \times \Delta\phi$ as longitudinal towers[3] with a total transverse energy above 2.5 GeV, used as seed for a sliding window algorithm.

Track seeds from the silicon detector are then extended to the calorimeter. First, the assumption that they are pions is used to account for the energy loss due to interactions with the detector material. If the track seed cannot be successfully extended to a full track under the pion hypothesis and it falls into an energy cluster in the EM, it is refitted under the hypothesis that it comes from an electron, which allows for larger energy losses. The final tracks are matched to EM clusters using the ΔR metric.

Clusters are then rebuilt using 3×7 (5×5) longitudinal towers of cells in the barrel (end-caps) of the EM calorimeter, with the energy of the clusters calibrated using studies based on simulated MC samples [98]. If no tracks are associated with an ECAL cluster, it is classified as a photon.

The final electron energy is given by the final calibrated cluster, while the η and ϕ directions are taken from the corresponding track parameters. Furthermore, electron measurements are performed by requiring the track to be compatible with the primary vertex, in order to reduce the background from conversions and decays of secondary particles.

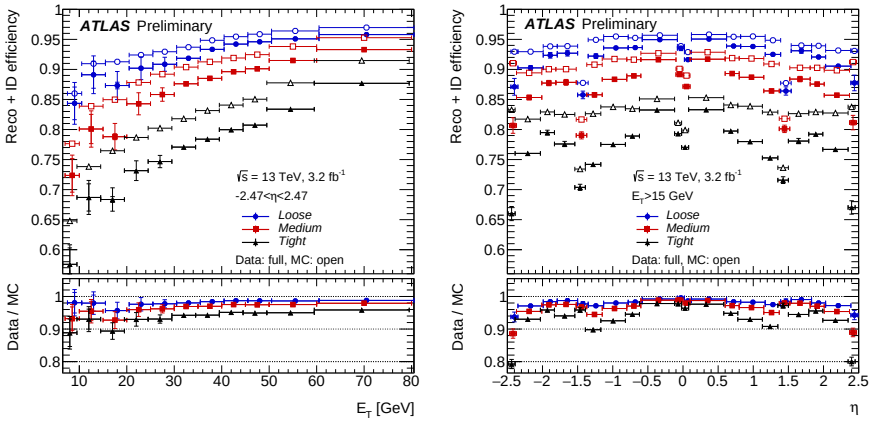
Lastly, for most of the analyses in ATLAS, including the ones presented in this thesis, electrons within the transition region between the barrel and end-cap of the calorimeter, $1.37 < |\eta| < 1.52$, are vetoed, as this region has a poor reconstruction and energy resolution performance.

Algorithms for electron identification (ID) are used in order to determine whether the reconstructed electron is indeed an electron or another object faking it, such as converted photons or jets. These algorithms use quantities related to the shape of the electromagnetic cluster shower, track properties, as well as track-cluster matching variables.

The baseline identification algorithm is a likelihood-based method, that allows for a simultaneous evaluation of several properties via the means of signal and background probability density functions (PDFs)

[3] A tower corresponds to the segmentation of ECAL in all the layers of the calorimeter.

of the discriminating variables. This allows for better background rejection for a given signal efficiency than a "cut-based" algorithm. Typically, three identification operating points are provided for electron ID, referred to as *Loose*, *Medium* and *Tight*, in order of increasing background rejection. Figure 3.4 shows the combined electron reconstruction and ID efficiency as a function of E_{T} and η.



(a) **(b)**

Figure 3.4: Combined electron reconstruction and identification efficiencies in $Z \to ee$ events as a function of the transverse energy E_{T} (a) and as a function of pseudorapidity η (b). These plots are taken from Ref. [97].

Another requirement imposed on reconstructed electrons is that they must be isolated, i.e. the detector activity around the electron must be minimal, in order to disentangle prompt electrons, such as those from $Z \to ee$ events, from other sources, like electrons originating from photon conversions, electrons from heavy flavour hadron decays and light hadrons (mostly pions) misidentified as electrons.

Two discriminating variables are used for this purpose:

- a calorimeteric variable, $E_{\mathrm{T}}^{\mathrm{cone0.2}}$, defined as the scalar sum of transverse energy clusters within a cone of $\Delta R = 0.2$ around the electron candidate, excluding the energy deposits associated with the candidate itself;

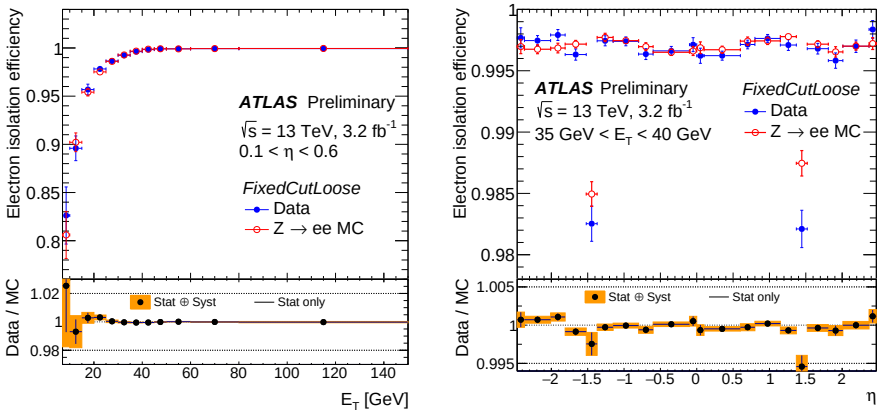
- a track-based variable, $p_{\mathrm{T}}^{\mathrm{varcone0.2}}$, defined as the scalar sum of the transverse momenta of all tracks within a cone around the candidate electron track of size $\Delta R = \min(0.2, 10\ \mathrm{GeV}/E_{\mathrm{T}})$ and stemming from the PV, excluding the electron associated tracks.

A variety of selection requirements on these quantities are defined, resulting in a set of eight isolation working points. More details on the characteristics of the various working points can be found in Section 5 of Ref. [97].

The accuracy with which the electron efficiency is modelled by detector simulation plays an important role in a variety of analyses, such as cross-section measurements and various searches for new physics. The efficiency to find and select an electron is not measured as a single quantity, but is divided into different components, namely reconstruction, identification, isolation, and trigger efficiencies, so that the total efficiency ε, is the product of the individual efficiencies, each one measured with respect to the previous step.

In order to achieve reliable physics results, the MC simulated samples need to be corrected in order to reproduce the measured efficiencies in data, therefore, a calibration of the MC detector response is needed. The calibration is provided in terms of multiplicative scale factors (SF) as a function of p_{T} and η of the electron, derived as the ratio of the efficiencies measured in data and the corresponding ones in simulation. The electron efficiencies are measured by using a tag-and-probe technique[4] using $Z \to ee$ events and $J/\psi \to ee$ events for the low p_{T} range. These data-to-MC correction factors are usually close to unity; deviations arise from the mismodelling of tracking properties or shower shapes in the calorimeter. Figure 3.5 shows the electron isolation efficiency as a function of the electron E_{T} and η for the representative *FixedCutLoose* isolation working point.

[4] The tag-and-probe method uses events containing resonances whose decay into particles is easy to identify, in this case $Z \to ee$ and $J/\psi \to ee$. A strict selection on one of the electron candidates (called "tag") together with the requirements on the di-electron invariant mass, and on the lifetime information for the case of J/ψ, allows for a loose pre-identification of the other electron candidate ("probe"). The probe electron is then used for the measurement of the reconstruction, identification, isolation and trigger efficiencies.



(a) **(b)**

Figure 3.5: Efficiency of the representative *FixedCutLoose* isolation requirement as a function of the transverse energy E_{T} (a) and as a function of η (b). Electrons are required to fulfil the *Tight* identification. For efficiencies as a function of E_{T} (η) the corresponding η (E_{T}) range used for the probe electron is shown on the plot. These plots are taken from Ref. [97].

3.2.2 Muons

Information from the inner detector and the muon spectrometer is primarily used for the muon reconstruction and identification, complemented by information from the calorimeters. Muon reconstruction is first performed independently in the ID and MS and then information is combined to reconstruct the final muon track, which is refitted using information from both systems. All figures and numbers in this section are taken from Ref. [99].

In the ID, muons are reconstructed in the same manner as all the other charged particles, whereas the muon tracks in the MS are built starting from track segments generated in the middle layers of the detector and then extended to use the segments from the outer and inner layers as seeds. At least two matching segments are required to build a track, except in the barrel-endcap transition region where a single segment can be used. In case the same segment is used to build more than one

track candidate, the best assignment to a single track is selected, unless the segment can be shared between two tracks, to ensure high efficiency for close-by muons. Finally, a new fit to obtain the final MS track is performed.

Four types of muons can be defined, depending on which sub-detector information was used in reconstruction:

Combined muons are built using information from both the ID and MS. A global fit is preformed to obtain the final muon track candidate, combining hits from both sub-detectors. This is the most common type of muon used for physics analyses.

Extrapolated muons are those for which the trajectory reconstruction is based only on the MS track, extrapolated to the interaction point. The estimated energy loss of the muon in the calorimeters is taken into account during the extrapolation.

Segment-tagged muons are reconstructed starting from an ID track extrapolated to match at least one track segment in the MS. They are used to increase acceptance for low-p_{T} muons or muons falling in regions with reduced MS acceptance.

Calorimeter-tagged muons have a track in the ID that can be matched to a calorimeter deposit compatible with a minimum-ionizing particle. This type of muon is used to recover acceptance in uninstrumented regions of the MS, used for cabling and services to the calorimeters and inner detector ($|\eta| < 0.1$), although it has the lowest purity among all muon types.

In order to identify prompt muons with high efficiency and reject background muons, such as muons originating from pion and kaon decays, identification requirements are implemented.

Muons originating from in-flight decays of charged hadrons in the ID are characterized by the presence of a "kink" in the reconstructed track. As a consequence, a poor fit quality of the resulting combined track is expected, accompanied by an imbalance of the momenta measured by the ID and MS alone. Such variables offer good discrimination

between prompt and background muons, and are employed, together with requirements on the number of hits, for muon identification, giving rise to four different identification selections: *Loose*, *Medium*, *Tight* and *High-*p_{T}, with the first three categories being inclusive categories, i.e. muons identified with tighter requirements are also included in the looser categories and the *High-*p_{T} selection aiming at maximizing the momentum resolution for tracks with p_{T} above 100 GeV, for searches for high-mass resonances. The *Medium* identification criteria provide the default selection for muons in ATLAS.

A tag-and-probe technique is used to measure the efficiency of the muon reconstruction and identification, within the acceptance of the ID, as well as to provide the necessary scale factors to account for mismodelling of the detector response in simulated samples. $Z \to \mu\mu$ events are used for the high-p_{T} part of the muon spectrum, while $J/\psi \to \mu\mu$ events are used to measure efficiencies up to 20 GeV. Figure 3.6 shows the reconstruction efficiency for *Medium* muons as a function of the p_{T} of the muon, in the region $0.1 < |\eta| < 2.5$.

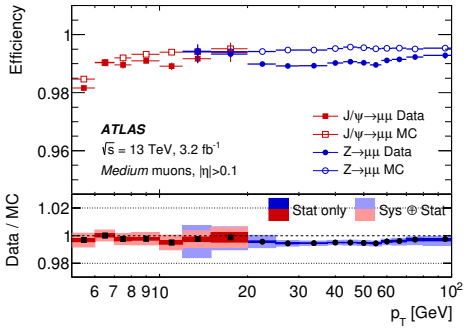


Figure 3.6: Reconstruction efficiency for the *Medium* muon selection as a function of the p_{T} of the muon, in the region $0.1 < |\eta| < 2.5$, as obtained with $Z \to \mu\mu$ and $J/\psi \to \mu\mu$ events. The error bars on the efficiencies indicate the statistical uncertainty. The panel at the bottom shows the scale factors for the muon efficiencies with both statistical and systematic uncertainties. From Ref. [99].

Analogously to the electron case, a measurement of the muon isolation is required in order to reject non-prompt or fake muons, such as

muons from light mesons or or semileptonic decays of b- or c-hadrons embedded in jets. Two variables measuring the detector activity around the muon are utilized for this purpose:

- a track-based variable, $p_{\mathrm{T}}^{\mathrm{varcone30}}$, defined as the scalar sum of the transverse momenta of tracks within a cone of variable size, $\Delta R = \min(0.3, 10\ \mathrm{GeV}/p_{\mathrm{T}})$, with $p_{\mathrm{T}} > 1$ GeV, around the muon candidate, excluding the muon track itself;

- a calorimeter-based variable, $E_{\mathrm{T}}^{\mathrm{topocone20}}$, defined as the sum of transverse energies of topological clusters within a cone of radius $\Delta R = 0.2$ around the muon candidate, excluding the energy deposits associated with the candidate itself. Corrections for pile-up and the underlying event are applied on an event-by-event basis.

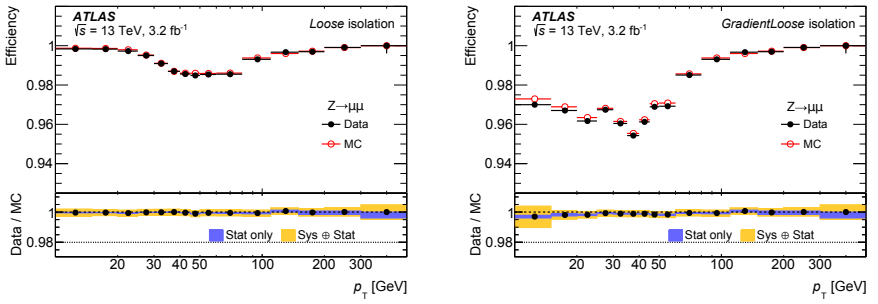
Seven isolation criteria are defined, based on a combination of the variables described above. All the isolation efficiencies are measured using $Z \to \mu\mu$ events, with a tag-and-probe method. Figure 3.7 shows the isolation efficiency measured for *Medium* muons in data and MC as a function of the muon p_{T} for the *Loose* and *GradientLoose* isolation working points.

The muon momentum scale and momentum resolution are studied using decays of J/ψ and Z to a di-muon pair. Even though the simulation contains an accurate description of the ATLAS detector, it is not enough to describe the muon momentum scale and momentum resolution to the desired level. To obtain a per mille level scale resolution and per cent level momentum resolution, a set of corrections is therefore applied to the simulated samples to match the resolution observed in data.

Independent corrections are derived for the ID and MS and then the momenta are combined to obtain the final corrected momentum.

Figure 3.8 shows the di-muon mass resolution divided by the invariant mass of the muon pair[5], as a function of the average p_{T} for J/ψ events and p_{T}^{*} for Z events. A detailed description of the method used, as well as a precise definition of the p_{T}^{*} variable, can be found in Section 8 of Ref. [99].

[5] If the two muons have similar momentum resolution and angular effects can be neglected, the relative mass resolution is directly proportional to the relative muon momentum resolution: $\sigma_{\mu\mu}/m_{\mu\mu} = \sigma_p/(\sqrt{2}p)$.



(a) **(b)**

Figure 3.7: Isolation efficiency as a function of the muon p_{T} for the *Loose* (a) and *GradientLoose* (b) muon isolation working points. The errors shown on the efficiency are statistical only. The bottom panel shows the ratio of the efficiency measured in data and simulation, as well as the statistical uncertainties and combination of statistical and systematic uncertainties. From Ref. [99].

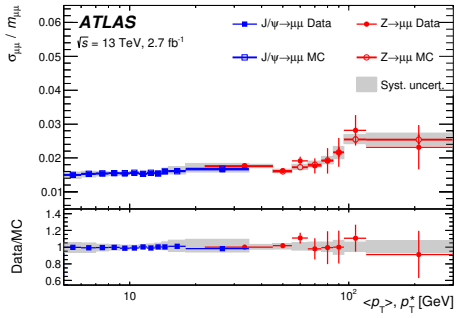


Figure 3.8: Di-muon invariant mass resolution divided by the di-muon invariant mass measured from $J/\psi \to \mu\mu$ and $Z \to \mu\mu$ events as a function of the average p_{T} for the J/ψ or p_{T}^{*} for Z events. For the precise definition of the p_{T}^{*}, the interest reader is referred to Eq. (12) in Section 8 of Ref. [99].

3.3 Jets

The time evolution of a quark or gluon produced in the final state of a collision predicts it to hadronize and fragment, creating a spray of collimated particles, reconstructed experimentally as a jet.

Calorimeter jets are reconstructed from topological energy clusters, starting with a seed cell in the calorimeters and adding iteratively neighbouring cells with an energy above the expected noise threshold [100]. Each topological clusters is calibrated, prior to jet reconstruction, to the electromagnetic scale (EM), which corresponds to the energy deposited by electromagnetically interacting particles in the calorimeter.

The most widespread jet clustering algorithm used in ATLAS is the anti-k_t algorithm [101]. It is a sequential algorithm: particles are clustered into jets one at the time, provided that their relative distance to other particles or the pseudo-jet, d_{ij}, is smaller than a stopping distance, d_{iB}. These distances are defined as:

$$\begin{aligned} d_{ij} &= \min(k_{ti}^{-2}, k_{tj}^{-2}) \frac{\Delta_{ij}^2}{R^2} \\ d_{iB} &= k_{ti}^{-2} \end{aligned} \tag{3.1}$$

where $\Delta_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2$ and k_{ti}, y_i and ϕ_i are the transverse momentum, the rapidity and the azimuthal angle of particle i, respectively. The distance d_{iB} is relative between the group of particles and the beam line; if it is the smallest distance the grouped particles form a jet. The typical radius parameter is $R = 0.4$.

This algorithm is such that soft particles will tend to cluster with hard ones before they cluster among themselves. If a hard particle has no hard neighbours within a distance $2R$, then it will simply accumulate all the soft particles within a circle of radius R, resulting in a perfectly conical jet.

The key feature is that the soft particles do not modify the shape of the jet, while hard particles do, i.e. the jet boundary in this algorithm is resilient with respect to soft radiation, but flexible with respect to hard radiation. Therefore the algorithm shows the desired theoretical properties of infra-red and collinear safety [101, 102].

Subsequently, reconstructed jets are calibrated to the jet energy scale (JES) derived from simulation and *in-situ* corrections based on 13 TeV data [103, 104]. The calibration proceeds in several steps, so that the final energy scale is that of *truth jets*[6]. Each stage of the calibration corrects the full four-momentum, scaling the jet p_{T} and energy. The flow of the calibration for EM-scale jets is shown schematically in Figure 3.9.

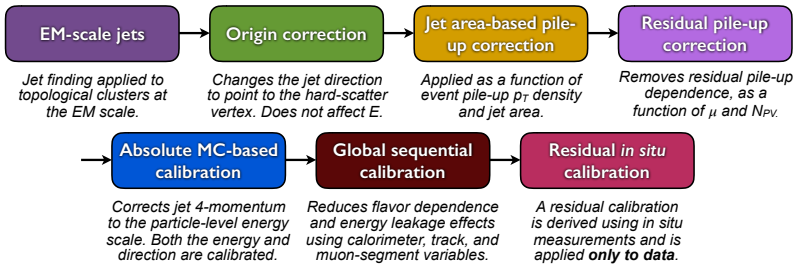


Figure 3.9: Calibration stages for EM-scale jets. Other than the origin correction, each stage of the calibration is applied to the jet four-momentum. From Ref. [103].

It consists of the following steps:

- origin correction: the four-momentum of jets is recomputed to point to the hard-scatter vertex rather than the centre of the detector, that is the default choice in the first step of the jet clustering. This step changes the direction of the jet keeping its energy constant, to improve the η resolution of jets.

- pile-up correction: this removes the excess energy due to in-time and out-of-time pile-up[7]. It consists of two components: an area-based pile-up contribution subtraction, as proposed in Ref. [105],

[6] A *truth-jet* is built with the same jet clustering algorithm, with the exception that it uses stable truth particles, i.e. particles with a mean lifetime $\tau > 3 \cdot 10^{-11}$s, that are able to travel through the detector before decaying, excluding muons and neutrinos.

[7] The in-time pile-up is due to multiple interactions happening in the same bunch crossing, whereas the out-of-time pile-up is due to the additional pp collisions happening in bunch-crossings just before and after the collision of interest. It is caused by the fact that the calorimeter integration time is longer than 25 ns.

applied at the per-event level; and a residual correction derived from MC simulations.

A correction is then applied, as a function of N_{PV} and the average number of interactions per bunch crossing, $\langle\mu\rangle$, to remove the residual dependence of the jet p_{T} on the pile-up.

- jet energy scale (JES) calibration: this is a correction that relates the reconstructed jet energy to the truth one. It is derived after the previous steps are applied, based on dijet MC events. It accounts for biases in the jet η reconstruction, mainly resulting from absorbed or undetected particles in gaps and transition regions between different sub-detectors.

- global sequential calibration: this aims at removing residual dependences of JES to the shower properties of the jet. The average particle composition and shower shape of a jet varies between quark- and gluon-initiated jets, as the former include, on average, hadrons with a higher fraction of the jet p_{T}, whereas a gluon-initiated jet tends to contain more, softer, particles. This step reduces the effects that lead to a different calorimeter response.

  During this step, a correction for the "punch-through" is carried out based on the amount of activity measured behind the jet, to correct high-p_{T} jets whose energy is not fully contained within the calorimeter jet. References [106, 107] contain an exhaustive description of the method.

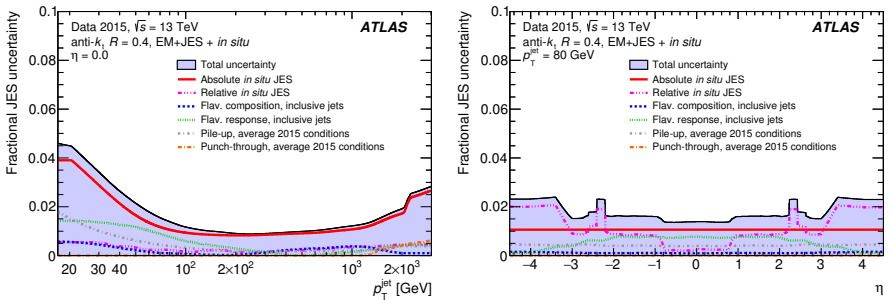
- *in-situ* calibration: this technique uses well-measured reference objects, including photons, Z bosons and calibrated jets and exploits imbalances among the reference objects.

  An η-intercalibration [104] corrects the average response of forward jets to match that of well-measured central jets using dijet events, whilst three other *in-situ* calibrations correct for differences in the average response of central jets with respect to those of well-measured reference objects, each one with a focus on a different p_{T} region. The correction is derived as a function of jet p_{T} and, in the η-intercalibration, also as a function of jet η.

For each *in-situ* calibration, the response is defined as the average ratio of jet p_{T} to reference object p_{T}. The combined response of the three methods is taken as the *in-situ* correction.

The full combination of all the various uncertainties described in the text above is shown in Figure 3.10 as a function of the jet p_{T} for jets at $\eta = 0$ and as a function of the jet η for jets with $p_{\mathrm{T}} = 80$ GeV.

The largest uncertainty is for low-p_{T} jets, starting at 4.5% and decreasing to 1% at 200 GeV, slightly increasing up to 2 TeV, when there is a sharp increase, due to the fact that the multi-jet balance measurement ends and the larger uncertainties are taken from the single-particle response studies. The uncertainty is fairly constant as a function of η and reaches a maximum of 2.5% for the most forward jets, except for a sharp feature which can be seen in the region $2 < |\eta| < 2.6$ due to the non-closure uncertainty of the η-intercalibration.



(a) **(b)**

Figure 3.10: Combined uncertainty in the JES of fully calibrated jets as a function of (a) jet p_{T} at $\eta = 0$ and (b) η at $p_{\mathrm{T}} = 80$ GeV. The various systematic uncertainty components are also shown. Figures taken from Ref. [103].

Jet cleaning

Several quality criteria are applied to identify jets arising from non pp collision sources, such as beam-gas interactions, detector noise or cosmic rays, using variables based on the shape of the signal in the ca-

lorimeters, energy thresholds and track-based variables [108]. Events containing such jets are discarded from further analysis.

In order to reduce in-time pile-up effects, an additional requirement on the output of the Jet Vertex Tagger (JVT) [109] is applied for jets with $p_{\mathrm{T}} < 60$ GeV and $|\eta| < 2.4$. This tagger uses a two-dimensional likelihood exploiting the differences between pile-up jets and jets originating from the hard-scatter. The discriminating variables are based on the p_{T} of the tracks associated with the jet, the jet p_{T}, as well as the scalar sum of the p_{T} of all the associated tracks originating from any of the pile-up interactions, corrected for the number of reconstructed primary vertices.

3.3.1 *b*-jets

The expression *b*-tagging is commonly used to indicate the identification of jets originating from the fragmentation of *b*-quarks, referred to as *b*-jets. It plays a vital role for precise Standard Model measurements, including the main focus of this thesis, $t\bar{t}H(b\bar{b})$, and for searching for New Physics signals, due to the fact that many New Physics scenarios have enhanced production of *b*-jets.

During hadronization, *b*-quarks form *b*-hadrons, which in the end decay via the electroweak interaction and thus a *b*-jet contains charged tracks coming from the decay of the *b*-hadron and tracks produced in the *b*-parton showering. Given their mean lifetime of the order of ps, *b*-hadrons travel a few millimetres from the primary vertex before decaying, resulting in a displaced decay vertex within a jet, the secondary vertex (SV). Furthermore, the main decay mode of a *b*-hadron is with the transition of a *b*- to a *c*-quark and the subsequent *c*-hadron presents a decay vertex displacement, even though smaller compared to the one of the *b*-hadron decay, resulting in a ter-

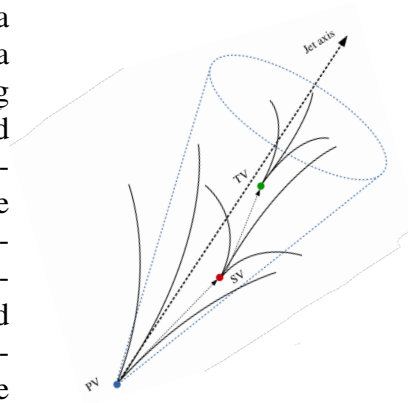


Figure 3.11: Drawing of the decay chain inside a *b*-jet.

tiary vertex (TV). The whole decay chain is depicted in Figure 3.11.

For simulated jets, a flavour label is assigned by matching jets to a truth-level weakly decaying b- or c-hadron within a cone of $\Delta R = 0.3$. The flavour labelling is exclusive, i.e. if a b-hadron is found within the cone the jet is labelled as a b-jet, otherwise in case no b-hadron is found, the search is repeated for c-hadrons and then for τ leptons. If no match is found, the jet is labelled as a light-jet.

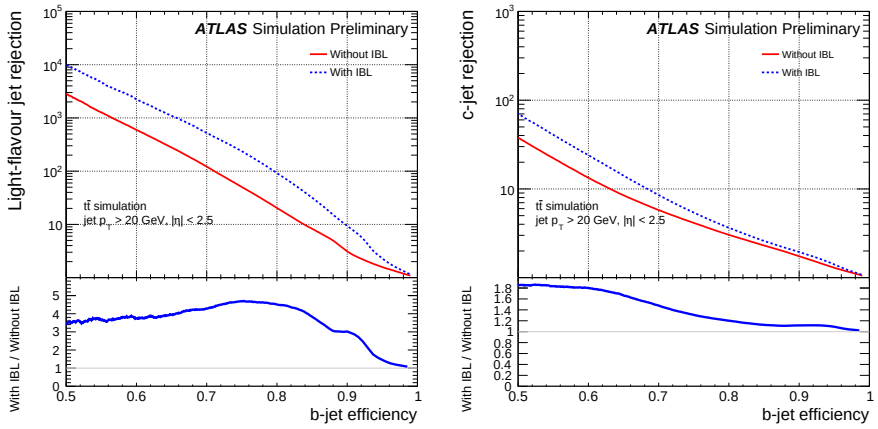
The Secondary Vertex Finder algorithm (SVF) implemented in the ATLAS software [110] proceeds to the reconstruction of a multi-track displaced vertex by first creating all the two-track vertex pairs with the given set of candidate tracks. Vertices that are thought to come not from the in-flight decay of heavy hadrons, such as vertices due to hadronic interaction with the detector material or K_s or Λ decays are rejected. Finally, a single multi-track vertex is reconstructed with the set of cleaned vertices that survived, with outlier tracks iteratively removed.

Secondary vertex reconstruction and b-tagging in general benefited vastly from the inclusion of the IBL before the start of Run2. This can be clearly seen in Figure 3.12, where the light- and c-jets rejections, i.e. the reciprocal of the efficiency, are shown as a function of the b-tagging efficiency.

Various quantities can be exploited to identify a b-jet experimentally, based on the properties of tracks and vertices of the jet, which leads to three general classes of algorithms: the impact parameter based algorithms, the inclusive secondary vertex ones and eventually the decay chain multi-vertex reconstruction algorithm. Several basic algorithms have been used in ATLAS, that fall into one of the three categories defined above:

IP2D, IP3D make use of both the longitudinal and transverse signed[8] impact parameter significances, $d_0/\sigma(d_0)$ and $z_0/\sigma(z_0)$, of the tracks matched to the jet. The probabilities under the b- and light-flavour jet hypotheses are evaluated using the PDFs of the impact parameter significance and later combined into a single likelihood ratio discriminant. IP3D uses both the transverse and longitudinal impact parameters taking into account their correlations, while

[8] The sign is defined positive (negative) if the point of closest approach of the track to the primary vertex is in front (behind) the primary vertex, relatively to the jet direction.



(a) **(b)**

Figure 3.12: Comparison of the performance of the *b*-tagging algorithm expressed in terms of light-jet (left) and *c*-jet rejection (right) as a function of *b*-tagging efficiency for the Run1 ("Without IBL") and Run2 ("With IBL") detector layouts. The underlying algorithms are updated to the detector geometry in each case. Plots taken from Ref. [111].

IP2D only uses the transverse impact parameters, making it more robust against pile-up effects, as $z_0/\sigma(z_0)$ is typically larger for tracks from such jets.

SV algorithms exploit properties of the reconstructed SV. Additional cleaning cuts are applied in order to reject fake vertices and then discriminating variables, such as the invariant mass of the vertex, its decay length significance or the number of tracks, are combined together into a likelihood.

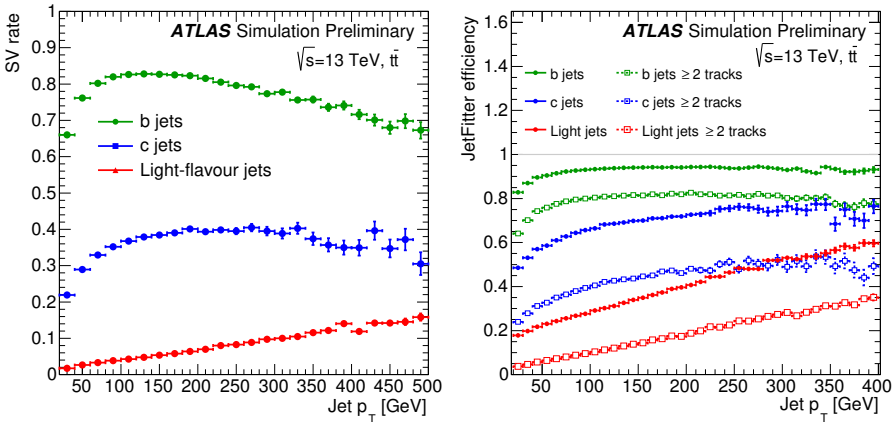
JetFitter [112] is the only multi-vertex reconstruction algorithm run in ATLAS. It tries to reconstruct the full PV $\rightarrow b \rightarrow c$-hadron decay chain by exploiting its topological structure.

One of its main assumptions is that *b*- and *c*-hadrons decay vertices lie on the same line defined by the *b*-hadron flight axis and that it corresponds to the jet axis, thus all charged particles stem-

ming from b- or c-decays intersect it. With this approach, the b- and c-hadron vertices can be resolved, even in the case of a single track attached to each of them.

As shown in Figure 3.13, the efficiency to reconstruct at least a single-track vertex is significantly higher than the efficiency to reconstruct a vertex with at least two tracks, but this comes at the price of a higher rate of reconstructed vertices in light-jets compared to the SVF reconstruction algorithm.

After JetFitter reconstruction, new variables can be constructed based on either the decay topology or the properties of the reconstructed vertices, such as the vertex decay length significances, the mass of the vertices and the energy fraction of the charged particles in the displaced vertices.



(a) SVF **(b)** JetFitter

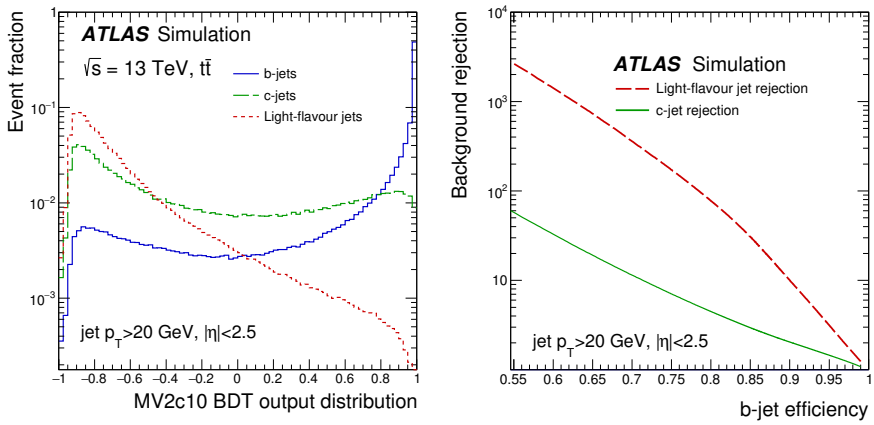
Figure 3.13: Secondary vertex reconstruction efficiency as function of jet p_{T} for the SVF (left) and JetFitter (right) for b- (green), c- (blue) and light-jets (red) evaluated using a sample of $t\bar{t}$ simulated events. The solid lines with closed markers represent the efficiency to reconstruct any JetFitter decay chain, the dashed line with open markers requires that at least one vertex has two or more tracks. Both figures are from Ref. [113].

The outputs of these basic b-tagging algorithms are later combined

in a multivariate discriminant, in order to provide the best separation among the three different jet flavours [113]. The Run2 algorithm constitutes a significant improvement over the main b-tagging algorithm used during Run1, as detailed in Ref. [114].

Various trainings have been performed by having as background different mixtures of light- and c-jets, resulting in different output discriminants named MV2cXX, where XX is the approximate fraction of c-jets used as background in the training.

In Figure 3.14a the output of the MV2c10 discriminant is shown, which is the default in ATLAS as it gives a good trade-off between light- and c-jet rejection for several analysis needs. The performance of the MV2 b-tagging algorithms is shown in Figure 3.14b for the light and c-jet rejection as a function of the b-jet efficiency. Both plots are evaluated using simulated $t\bar{t}$ events, considering jets with $p_{\mathrm{T}} > 20$ GeV and $|\eta| < 2.5$.



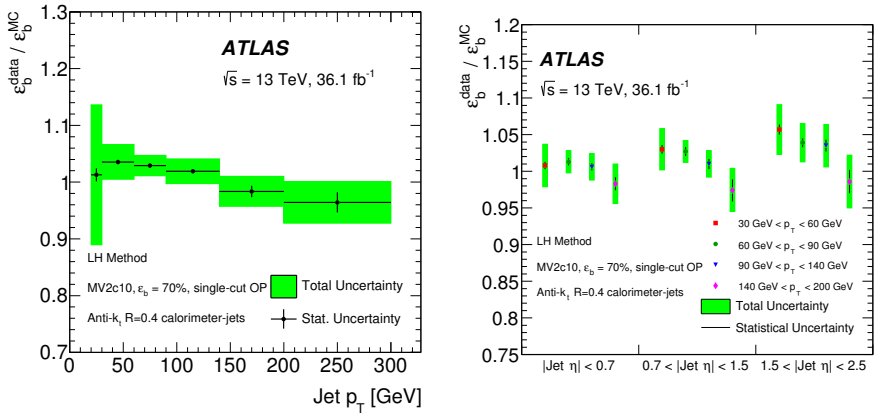
(a) **(b)**

Figure 3.14: (Left) the MV2c10 output for b-jets (solid blue), c-jets (dashed green) and light-flavour jets (dotted red) and (right) the light-flavour jet (dashed red) and c-jet (solid green) rejection factors versus b-jet tagging efficiency of the MV2c10 b-tagging algorithm, evaluated on $t\bar{t}$ simulated events. Plots taken from Ref. [115].

Four operating points have been defined for the recommended algo-

rithm, defined by the b-jet efficiencies of 85%, 77%, 70% and 60%. These efficiencies have been calibrated in data, using samples enriched with b-, c- or light-jets, and the result of the calibrations is presented in terms of efficiency SF, defined as SF=$\varepsilon_{\text{data}}/\varepsilon_{\text{MC}}$.

The measurement of the b-tagging SF comes from a Combinatorial Likelihood Method (LH), exploiting the presence of b-jets produced in $t\bar{t}$ dileptonic events [115]. Scale factors as a function of probe jet p_{T} and $|\eta|$ for b-jets are shown in Figure 3.15, for the representative 70% working point. The fact that they are very close to unity is indicative of the good modelling, in simulation, of the final MV2c10 discriminant.



(a) **(b)**

Figure 3.15: Data-to-MC scale factors corresponding to the 70% b-jet tagging efficiency as a function of the jet p_{T} (a) and as a function of the jet $|\eta|$ (b), obtained using the LH method as described in Ref. [115]. Both the statistical uncertainties (error bars) and total uncertainties (shaded region) are shown.

3.4 Missing transverse energy

Weakly interacting particles, like neutrinos or particles predicted by Beyond the Standard Model theories, escape the ATLAS detector unde-

tected, hence creating an imbalance in the measured sum of the momentum in the transverse plane.

The missing transverse momentum, $E_{\mathrm{T}}^{\mathrm{miss}}$, is an important observable used to measure the transverse momentum carried by such undetected particles. The reconstructed $E_{\mathrm{T}}^{\mathrm{miss}}$ is characterized by two contributions: the first contribution comes from the hard-event (calibrated) objects and the second contribution comes from the *soft term* signal, consisting of reconstructed charged-particle tracks not associated with the accepted hard objects, but still associated with the primary vertex of the event [116]. In mathematical language:

$$E_{x(y)}^{\mathrm{miss}} = -\sum_{i \in \text{ hard objects}} p_{x(y),i} - \sum_{i \in \text{ soft signals}} p_{x(y),i} \tag{3.2}$$

In the calculation of the quantities in Eq. (3.2) the contributing objects need to be reconstructed from mutually exclusive detector signals, in order to avoid double-counting of the same signal in more than one reconstructed observable. This is done by considering physics objects in a specific order, the most common sequence is considering first electrons, followed by photons, hadronically decaying τ-leptons and finally jets. Muons have little to no signal overlap with the other reconstructed particles in the calorimeter. If a low-priority object (γ, hadronically decaying τ lepton) shares the calorimeter signal with a higher-order priority object, it is fully rejected.

Another important challenge to the proper computation of $E_{\mathrm{T}}^{\mathrm{miss}}$ is represented by pile-up. The most important contribution comes from the additional particles created in the extra collisions, but a second, relevant, contribution comes from the out-of-time pile-up due to neighbouring bunch crossings.

Pile-up is fought by using calibrated jets, with $p_{\mathrm{T}} > 20\mathrm{GeV}$, by using the jet cleaning techniques described above, and by using a track-based computation of the soft term, to make it more robust against pile-up.

Uncertainties affecting the hard objects are directly propagated to the final computation of $E_{\mathrm{T}}^{\mathrm{miss}}$, therefore they are not considered as a separate source of systematic uncertainty for its computation. On the other hand, variations in the computation of the soft term give rise to separate, genuine uncertainties.

Different MC samples have differences in the soft term among themselves larger than what is observed in data, therefore the systematic uncertainties are evaluated based on different generated MC samples. Systematic uncertainties are assigned to the longitudinal and transverse resolution with respect to $p_{\mathrm{T}}^{\mathrm{hard}}$, defined as the vectorial sum of the p_{T} of the hard objects in the event, as well as on the scale of the soft term itself. The impact of these systematics is negligible in the analyses discussed in this thesis, as the missing transverse momentum is not used for event selection, but only in the event reconstruction.

4 Jet Vertex Charge

Quarks are confined inside hadrons, due to the hadronization process discussed in Section 1.1.1, hence their properties, such as the electric or colour charge, cannot be directly accessed. Nevertheless, indirect charge measurements are possible and first attempts in this direction date back to the introduction of a quark-charge sensitive observable first suggested by Field and Feynman [117]. The proposal was to construct a variable sensitive to the quark electric charge, a *jet charge*, Q_J, by the means of a sum of all the charges of the tracks present in a jet with a decreasing weight, in order to suppress fluctuations from extra or missing tracks and maximize the sensitivity to leading particles, which tend to carry most of the information from the fragmentation process.

The jet charge observable has been investigated extensively in ATLAS in dijet, W+jets and semileptonic $t\bar{t}$ events using 8 TeV data [118–120]. Further implementations of these ideas, in different contexts, have been studied within the ATLAS Collaboration, in particular, in the context of CP violation studies in the B_s system [121] and for the measurement of the top quark charge in pp collisions at $\sqrt{s} = 7$ TeV [122]. In addition, measurements using Vertex and Kaon Charge tags were done at the SLAC Detector, which used the net charge of the displaced vertices, as well as the charge of tracks stemming from vertices that are identified as kaons [123].

The *Jet Vertex Charge* tagger (JVC) is an evolution of these ideas with the precise aim of improving the performance for b-jets [124]. Final states with a high number of jets and b-jets are numerous in the SM and given the large combinatorics, it is hard to understand which jet comes from which particle: the possibility to distinguish between jets originating from b-quarks and $\bar{b}$-quarks can thus provide useful information for reducing this background and for helping a full final state

reconstruction.

Among those final states there is the one investigated in this thesis, the Standard Model Higgs boson produced in association with top quarks. In this analysis, the overwhelming $t\bar{t}$ + jets background precludes the possibility to directly identify and reconstruct the Higgs boson; reducing the number of possible associations can provide an important boost in the analysis.

This chapter presents the description of both the Jet Vertex Charge algorithm itself and its calibration analysis, done using 2015 and 2016 data collected by the ATLAS detector at $\sqrt{s} = 13$ TeV [125].

4.1 The tagger

All the different taggers that aim at identifying b-jets exploit, as described in Section 3.3.1, the distinctive signs of the b-hadron decay: the presence of displaced vertices.

JVC is no different in this respect, as it exploits extensively the topology and kinematics of the b-hadron decay chain reconstructed by the JetFitter algorithm [112]. The advantage of using this algorithm resides in the fact that not only it is the only vertex finder algorithm run in the ATLAS software that is able to reconstruct tertiary vertices, but it is also able to reconstruct single-track displaced vertices.

Additional information exploited by JVC are charge variables associated with the decay vertices, the so-called secondary (Q_{SV}) and tertiary (Q_{TV}) jet charge. Above that, semileptonic decays into a muon provide important information, which is also exploited by the algorithm.

Eventually, all the different information is combined by means of a multivariate analysis (MVA) in order to obtain a final discriminant, λ_{JVC}, which can be interpreted as the ratio of the likelihoods for a b-jet to have a positive or negative charge. In this way it is therefore possible to tag a jet as coming from a positively or negatively charged b-jet.

4.1.1 Algorithm

The tagger has been trained and optimized using jets with a p_{T} greater than 20 GeV and $|\eta| < 2.5$.

The truth flavour of the jet is assigned following the same cone-labelling procedure described in Section 3.3.1 on page 78: the flavour label is assigned by matching jets to a truth-level weakly decaying b- or c-hadron within a cone of $\Delta R = 0.3$.

The *truth charge* of the b-jet is then assigned by looking at the quark content of the hadron used to tag the jet: hadrons containing a $\bar{b}$ (b) quark are assigned a positive (negative) charge. In case of multiple b-hadrons matched with the jet, the one with the highest p_{T} is selected.

The phenomenon of neutral b-meson oscillations complicates the picture just described, as one needs to further specify the hadron used to assign the truth charge to a given jet: the hadron produced via the strong interaction, prior to any oscillation, or the weakly decaying one, after possible oscillations have occurred.

For the purpose of the MVA training, the b-hadron used to define the truth charge of the jet is the weakly decaying one. This choice is motivated by the fact that the algorithm showed the best performance because, in this way, there is a stronger correlation between the truth flavour definition and the input variables of the algorithm. On the other hand, in the evaluation of the performance and in the calibration analysis, this effect is properly taken into account, as explained in Section 4.1.5.

4.1.2 Jet Charge Variables

The basic variable that can be constructed to measure the charge of a jet is given by a p_{T}-*weighted* sum of the charges of the tracks associated with the jet:

$$Q_{\mathrm{J}} = \frac{\sum_{i \in \mathrm{Trk}} q_i \cdot {p_{\mathrm{T}}}_i^{\kappa}}{\sum_{i \in \mathrm{Trk}} {p_{\mathrm{T}}}_i^{\kappa}} \tag{4.1}$$

where the index i runs over the set of tracks associated with jet with electric charge q_i and transverse momentum $p_{\mathrm{T}i}$. The free parameter κ is used to maximize the separation power between the distribution of positive and negative b-jets. In the denominator the simple sum of the p_{T} of the tracks is preferred to other normalizations, such as the jet

transverse momentum, in order to avoid relying on calorimetric information.

Different definitions of the jet charge can be constructed using different reweighting schemes. A straightforward variation is the use of the $p_{\mathrm{T}}^{\mathrm{Rel}}$ or the $p_{\mathrm{T}}^{\mathrm{Long}}$, the component of the momentum transverse or along the jet axis, instead of the simple p_{T}. Both definitions have been tested with no noticeable difference, therefore the simplest solution has been adopted.

The optimal values of the κ parameter were found to be $\kappa = 1.1$ for the Q_{J} and $\kappa = 0.7$ for both Q_{SV} and Q_{TV} variables.

Basic selection cuts on the tracks are applied in order to capture tracks coming from the decay of the b-hadron and ensure optimal separation. The tracks used to calculate the Q_{J} variable are required to satisfy $p_{\mathrm{T}} > 500$ MeV, $|\eta| < 2.5$ and $\chi^2/N_{\mathrm{df}} < 5$. In addition, their transverse and longitudinal impact parameters must satisfy $|d_0| \leq 3.5$ mm and $|z_0 \sin\theta| \leq 4.5$ mm. Finally, a minimum of one hit in the pixel detector, four hits in the SCT, with a minimum of nine hits summing both the pixel and the SCT, and nine in the TRT are required, with at most one shared hit between different tracks.

While the JetFitter approach offers important advantages, it leads to a significant rate of fake vertices. The number of reconstructed vertices can be easily greater than the two expected. In order to remove fake vertices, additional quality cuts are applied to ensure a good fake vertex rejection. They include a cut on the fitted $\chi^2/N_{\mathrm{df}} < 5$ and on the error of the reconstructed flight length ($\Delta L_{3\mathrm{D}}$) less than 5 mm. Furthermore, vertices with fitted flight lengths $L_{3\mathrm{D}} > 250$ mm are discarded to improve the separation power of the Q_{SV} variable while taking as many SVs into account as possible. The remaining vertices are ordered according to the distance with respect to the PV: the closest vertex is assumed to be the SV, also in the case it is reconstructed as a single-track vertex, whereas all the other displaced vertices are combined into a single vertex, identified as the TV. It is found that for single-track vertices, the observed charge corresponds to the expected one in about 65% of the cases, which is the same fraction found for multi-track SV, providing evidence for the hypothesis that single-track SVs are not significantly affected by fake tracks.

In Figure 4.1 the distributions for the three jet charge variables are shown for *truth b-jets* with $p_{\mathrm{T}} > 20$ GeV and $|\eta| < 2.5$, which is the jet kinematic selection applied in the following, unless stated otherwise. In addition to the charges described above, the distribution of the jet charge without any selection cuts, $Q_{\mathrm{J}}^{\text{all tracks}}$, is shown. This charge is employed in the tagger only in the case no other information is available. The basic charge variables Q_{J}, Q_{SV} and Q_{TV}, show regular distributions in the interval $(-1,+1)$ with a broad peak around 0. These smooth shapes are accompanied by spikes at the values of ± 1, mainly populated by the jets in which the charge in question is computed either using a single track or a few tracks with the same charge.

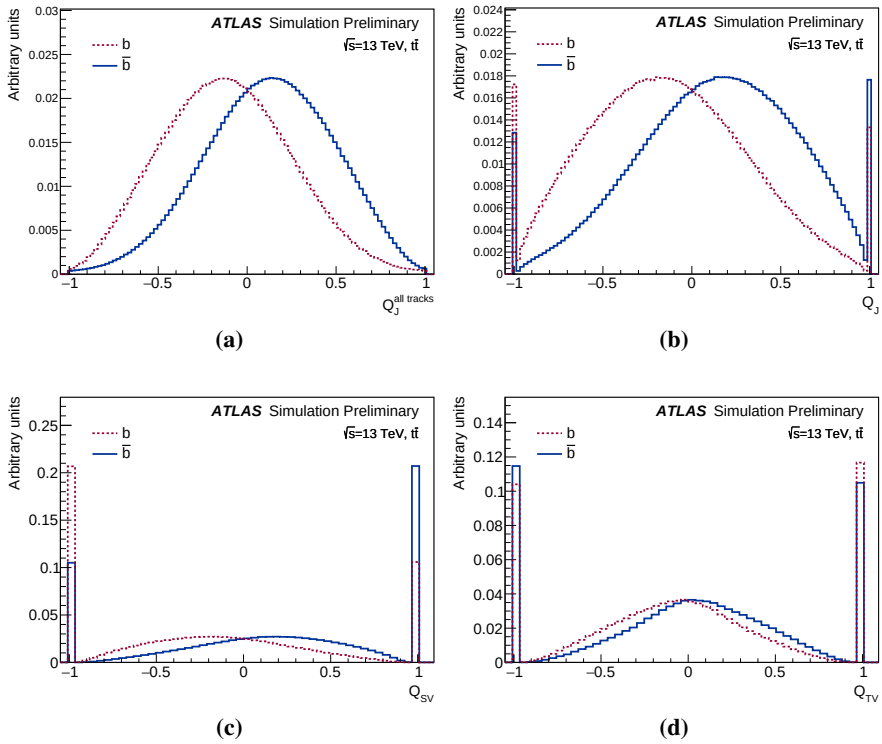


Figure 4.1: Distributions of the basic charge variables for positive (blue) and negative (red) truth *b*-jets: (a) jet charge computed from all available tracks ($Q_{\mathrm{J}}^{\text{all tracks}}$), (b) jet charge ($Q_{\mathrm{J}}$), (c) secondary vertex charge (Q_{SV}) and (d) tertiary vertex charge (Q_{TV}).

A final remark must be made about the set of tracks used to compute Q_{TV}; in fact, if an odd number of tracks are associated with the TV, the lowest p_{T} track is ignored. This choice is motivated by the fact that, according to Monte Carlo studies, the majority of c-hadrons that decay into the TV are $D^0/\overline{D^0}$ [1] and the decay topology of these mesons gives always an even number of prongs [8]. The improvement given by this further track selection can be seen in Figure 4.2.

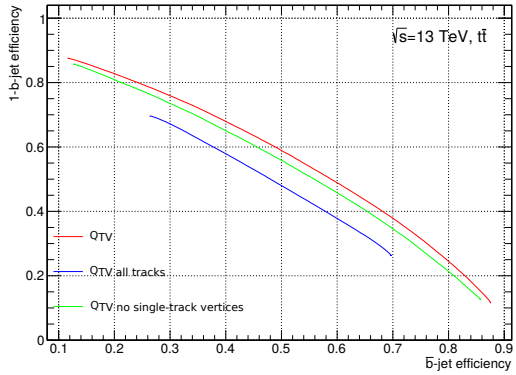


Figure 4.2: Comparison of the separation power between positive and negative b-jets achieved with three different definitions of the tertiary vertex charge: Q_{TV} computed considering all tracks associated with the TV (blue line), considering all tracks, but rejecting single-track vertices (green line) and computed discarding the lowest p_{T} track in case of an odd number of tracks (red line).

4.1.3 Soft Muon Charge

Semileptonic b-hadron decays are a source of valuable information, as the charge of the lepton carries the same sign of the charge of the underlying b-quark contained in the weakly decaying b-hadron, as can be

[1] The fraction of $D^0/\overline{D^0}$ mesons corresponds to 55.2% of the total, followed by $D^{\pm}$ (23.9 %), $D_s^{\pm}$ (14.0 %), $\Lambda_c^{\pm}$ (4.3%) and other mesons contribute to the remaining 2.6 %.

seen in Figure 4.3. On the contrary, if the muon comes from the subsequent *c*-hadron decay it carries the opposite charge of the *b*-hadron.

Only *combined* muons, which are muons with a reconstructed track both in the Inner Detector and in the Muon Spectrometer are considered, provided they are within a cone of radius $\Delta R = 0.3$ around the jet. Selection cuts involve the quality of muon reconstruction, such as $\chi^2/N_{\mathrm{df}} < 5$ of the match between the track reconstructed in the ID and the MS, as well as basic kinematic cuts of $p_{\mathrm{T}} > 5$ GeV and $|\eta| < 2.5$. In order to avoid using isolated muons, the sum of the p_{T} of tracks within a cone of radius 0.3 around the muon must be $I_\mu = \Sigma p_{\mathrm{T}}/p_{\mathrm{T}}^{\mu} > 0.05$. If more than one muon is associated with the jet, only the muon with the highest p_{T} is considered. The charge of the considered muon, Q_μ, is shown in Figure 4.4, where no distinction is made between the muons originating from *b*- and *c*-hadrons, which results in the similar rates of the muons with the same and opposite charge and a dilution of the performance of the variable.

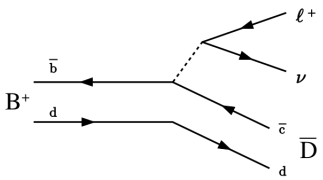


Figure 4.3: Feynman diagram of a semileptonic *b*-meson decay: the charge of the meson is reflected in the charge of the lepton.

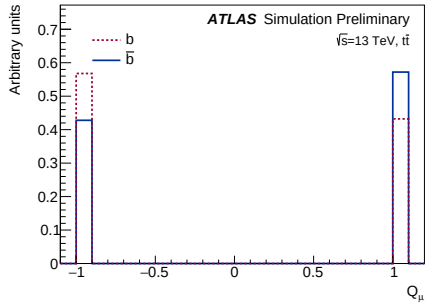
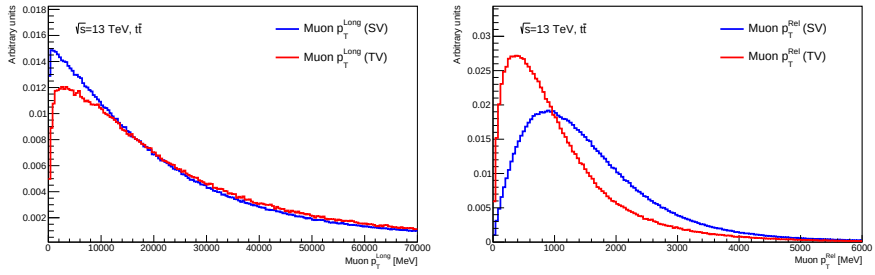


Figure 4.4: Distributions of the muon charge variable for positive (blue) and negative (red) truth *b*-jets.

Understanding the origin of the muon is a crucial point in order to extract the maximal information from this variable, therefore additional variables are employed in order to identify the decay vertex that it originated from. The $p_{\mathrm{T}}^{\mathrm{Rel}}$ and $p_{\mathrm{T}}^{\mathrm{Long}}$ of the muon are used for this purpose, which represent the momentum of the muon perpendicular to and along the jet plus muon axis, respectively. As is shown in Figure 4.5, at truth

level muons originating from *b*-hadron decays have a harder spectrum than muons originating from *c*-hadron decays.



(a) **(b)**

Figure 4.5: $p_{\mathrm{T}}^{\mathrm{Long}}$ (right) and $p_{\mathrm{T}}^{\mathrm{Rel}}$ (left) of muons coming from a SV (blue) or TV (red) at truth level.

4.1.4 Multivariate Analysis

The usage of multivariate analysis techniques in high energy physics has become more and more important in recent years. They are of particular importance in situations when it is not possible to identify a single variable that exhibits a clear separation between signal and background to cut on.

These techniques are based on the so called *Machine Learning*[2], a branch of the artificial intelligence that focuses on building systems that are able to learn from the data they are exposed to.

MVA techniques provide a valuable way to solve the problem of event classification. As opposed to a traditional cut-based selection,

[2] The term "Machine Learning" was first coined in 1959 by Arthur Samuel, a researcher at IBM, while successfully designing a self-learning program able to play checkers [126]. Nowadays it refers to the field of computer science that tries to give to computers the ability to learn and identify a specific pattern or behaviour without being taught so. As time progressed, more and more sophisticated versions of self-learning algorithms were developed. The most notable examples are the IBM computer Deep Blue, which was the first computer able to win, in 1996, a chess game against Kasparov, the world champion at that time, and the more recent AlphaGo, "the first computer program to defeat a professional human Go player, the first program to defeat a Go world champion, and arguably the strongest Go player in history" [127, 128].

they make use of a multi-dimensional observable space rather than each observable separately, being able in this way to extract the maximum information and improve signal over background discrimination. The classification of each event in a particular class, being either signal or background, is based upon the use of distributions of the events in the phase space described by the chosen observables: they combine a vector of variables describing the event, $\boldsymbol{x}$, into a single variable, y. It can be thought as a map from a D-dimensional space to $\mathbb{R}$, $\mathbb{R}^D \to \mathbb{R} : y = y(\boldsymbol{x})$. Simplifying, each constant value $y = c$ represents a hypersurface in the $\mathbb{R}^D$ space and classifying events with $y > c$ is equivalent to labelling all events on one side of the hypersurface, however complicated it is, as signal and rejecting all the others as background. Rather than labelling an event as belonging to a definite class, it is common to assign a value describing the likelihood of being of that particular class, often being in the interval [-1;1]. More information on the topic can be found in Refs. [129, 130].

The charge variables described previously are sensitive to the charge of the quark that initiated the jet, but individually do not provide optimal discrimination, hence MVA techniques are employed to better recognize the region of the phase space where the b- and $\bar{b}$-jets live.

To keep the analogy of the jet charge sign and the numerical values of the MVA discriminant, jets containing a weakly decaying $\bar{b}$ (b)-hadron are considered signal (background) in the MVA trainings. No b-tagging requirement has been imposed to select the sample of truth b-jets in order to keep this algorithm independent of any b-tagging algorithm and thus applicable in the combination with any of them.

The full categorization of the b-jets consists of eight exclusive groups, according to the availability of the basic charge variables, as summarized in Table 4.1. The category names consist in the list of the available jet charges, followed by the number of available muons.

In the categories labelled as $\mathrm{C_J}$, $\mathrm{C_{SV}}$ and $\mathrm{C_{all}}$, the discrimination between the b- and $\bar{b}$-initiated jets relies entirely on the only charge available in this category, i.e. Q_{J}, Q_{SV} and $Q_{\mathrm{J}}^{\text{all tracks}}$ variables respectively, given that no significant improvement has been found from training a dedicated MVA. In all other categories, available information is combined in the corresponding MVA discriminant trained using a Neural

Table 4.1: Availability of basic charge variables per category. The meaning of the symbols is: • variable is available, ○ variable is not available, – variable was not asked for.

Category	Q_J	Q_{SV}	Q_{TV}	Q_μ	$Q_J^{\text{all tracks}}$
C_J	•	○	○	○	–
$C_{J,\,\mu}$	•	○	○	•	–
C_{SV}	○	•	–	–	–
$C_{J,\,SV}$	•	•	○	○	–
$C_{J,\,SV,\,\mu}$	•	•	○	•	–
$C_{J,\,SV,\,TV}$	•	•	•	○	–
$C_{J,\,SV,\,TV,\,\mu}$	•	•	•	•	–
C_{all}	○	○	–	–	•

Network (NN), the Multi-layer Perceptron (MLP) method implemented in the TMVA toolkit [130].

Given that the basic variables show spikes in the distributions, to reduce the impact of those spikes, to smooth the shape of the MVA output and to improve the performance of the MVA, further variables are included in the machine learning procedure, called *auxiliary variables* in the following. Each basic charge variable is supported by its own set of corresponding auxiliary variables; for example, vertex reconstruction quality or the corresponding track multiplicity are intended to help the MVA to distinguish well reconstructed vertices from fake and poorly reconstructed ones.

Since a naïve association with the vertex is not sufficient for discriminating between muons coming from a *b*- or *c*-hadron decay, the muon is described by its kinematics, such as space angle between the muon direction and the jet axis, the p_T^{Rel} and p_T^{Long}, which help the MVA to differentiate between the same-sign and opposite-sign muon charge cases.

In each category, several MLP configurations have been tested using various sets of auxiliary variables. The final combination of variables and hyperparameters[3] was chosen to ensure smooth and symmet-

[3] The hyperparameters of an MVA are the adjustable parameters that determine the structure of the algorithm, such as the number of neurons or hidden layers in a NN.

ric shapes of the MVA output distributions populated by negative (b-initiated) and positive ($\bar{b}$-initiated) jets, as well as a low level of overtraining, i.e. the extreme specialization of the network which interprets statistical fluctuations as relevant information during the training stage. Typically, in categories with fewer basic charges available, more auxiliary variables are needed to provide a good quality of the training. A summary of all auxiliary variables along with a brief description can be found in Table 4.2, whereas the full list of the variables used in each category is presented in Table 4.3.

Finally, the best available discriminant is constructed for each category, referred to as the *JVC weight*, w. The discriminant relies on a single variable in three categories and on the corresponding MLP output in the other five categories. The JVC weight distributions normalized to unity are shown in Figure 4.6 for all eight categories. The plots show the overlaid distributions for the positive (blue line) and negative (red line) b-jets.

Combining the Categories

A given JVC weight w corresponds to different points in the space of the positive b-jet efficiency and the negative b-jet rejection depending on the category to which the jet in question belongs, due to the different shapes of the JVC weights across the categories. Thus, a more general discriminant with a unique interpretation across the full spectrum of its values is constructed: for each category, the JVC weight distributions for positive b-jets, $\bar{b}(w)$, and for negative b-jets, $b(w)$, are normalized to unity and, for a given weight w, the logarithm of the likelihood ratio:

$$\lambda_{\mathrm{JVC}}(w) = \ln \frac{\bar{b}(w)}{b(w)} \tag{4.2}$$

is used, as the variable offering the best discrimination between positive and negative b-jets, as prescribed by the Neyman-Pearson lemma [131]. The λ_{JVC} distribution obtained in this way for all the jet categories combined is presented in Figure 4.7.

Table 4.2: List of the auxiliary variables and their description.

Variable	Description
Q_{J}	See text for a detailed explanation.
$Q_{\mathrm{J}}^{\text{all tracks}}$	See text for a detailed explanation.
Q_{SV}	See text for a detailed explanation.
Q_{TV}	See text for a detailed explanation.
Q_{μ}	Charge of the muon associated with the jet.
$N_{\mathrm{trk}}(Q_{\mathrm{J}})$	Number of tracks used to compute Q_{J}.
$p_{\mathrm{T}}^{\mathrm{trk}}(Q_{\mathrm{J}})$	p_{T} of the hardest track used to compute Q_{J}.
$N_{\mathrm{trk}}(\mathrm{SV})$	Number of tracks in the SV.
$p_{\mathrm{T}}^{\mathrm{trk}}(\mathrm{SV})$	p_{T} of the hardest track used to compute Q_{SV}.
$L_{3D}(SV)$	Distance between the SV and the PV along the jet-axis.
$\Delta L_{3D}(SV)$	Error on the fitted position of the SV.
$m(\mathrm{SV})$	Invariant mass of the SV computed under the hypothesis that all the particles are pions.
$N_{\mathrm{trk}}(Q_{\mathrm{TV}})$	Number of tracks used to compute Q_{TV}.
$L_{3D}(\mathrm{TV})$	Distance between the TV and the PV along the jet-axis
$\Delta L_{3D}(\mathrm{TV})$	Error on the fitted position of the TV
$m(\mathrm{TV})$	Invariant mass of the TV, computed under the hypothesis that the most energetic particle is a kaon and the remaining particles are pions.
$p_{\mathrm{T}}^{\mathrm{Rel}}(\mu)$	Momentum of the muon orthogonal to the jet plus muon axis.
$p_{\mathrm{T}}^{\mathrm{Long}}(\mu)$	Momentum of the muon projected onto the jet plus muon axis.
$\Delta R(\mu,\mathrm{jet})$	ΔR angle between the muon and the jet axis.
$I_{40}^{\mathrm{var}p_{\mathrm{T}}}(\mu)$	Track momentum sum contained in the p_{T}-dependent cone of maximal size $\Delta R = 0.4$ around the muon.

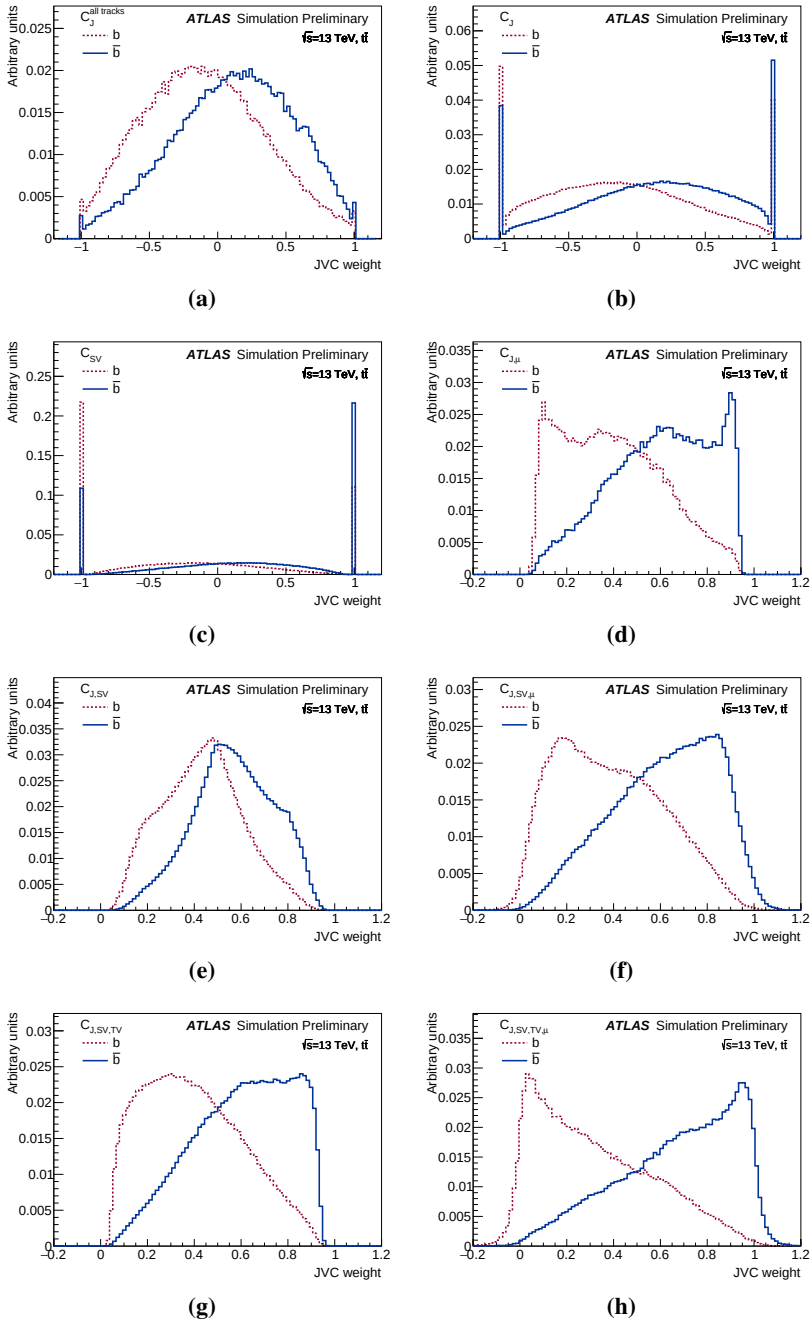


Figure 4.6: JVC weight distributions for b-jets (dashed lines) and $\bar{b}$-jets (solid lines) for the categories: C_{all} (a), C_J (b), C_{SV} (c), $C_{J,\mu}$ (d), $C_{J,SV}$ (e), $C_{J,SV,\mu}$ (f), $C_{J,SV,TV}$ (g), $C_{J,SV,TV,\mu}$ (h).

Table 4.3: List of input variables per MVA category. The definitions of the categories correspond to those in Table 4.1. For the explanation of the variable definitions see Table 4.2.

Variable	$C_{J,\mu}$	$C_{J,SV}$	$C_{J,SV,\mu}$	$C_{J,SV,TV}$	$C_{J,SV,TV,\mu}$
Q_J	•	•	•	•	•
$N_{trk}(Q_J)$	•				
$p_T^{trk}(Q_J)$	•	•	•	•	•
Q_{SV}		•	•	•	•
$N_{trk}(SV)$		•	•	•	•
$p_T^{trk}(SV)$		•		•	
$L_{3D}(SV)$		•	•	•	•
$\Delta L_{3D}(SV)$		•	•	•	•
$m(SV)$				•	
Q_{TV}				•	•
$N_{trk}(Q_{TV})$				•	•
$L_{3D}(TV)$				•	•
$\Delta L_{3D}(TV)$				•	•
$m(TV)$				•	•
Q_μ	•		•		•
$p_T^{Rel}(\mu)$	•		•		•
$p_T^{Long}(\mu)$	•		•		•
$\Delta R(\mu, jet)$	•				
$I_{40}^{var p_T}(\mu)$	•				

4.1.5 Performance

The performance of the final discriminant λ_{JVC} is evaluated in terms of the negative b-jets (background) rejection as a function of the positive $\bar{b}$-jets (signal) efficiency.

The weakly decaying b-hadron is used to define the truth charge of a jet for the MVA training; while this definition is useful because it relates the underlying physics of the decay to the input variables, using weakly decaying b-hadrons ignores the effects of b-meson mixing. Therefore, in order to take this effect into account in the evaluation of

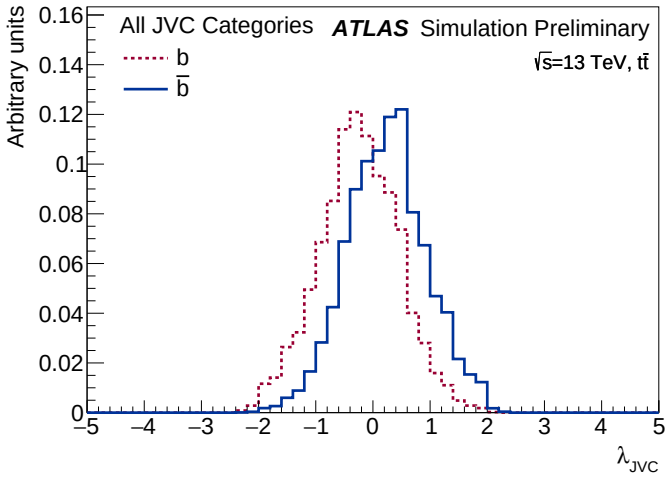


Figure 4.7: λ_{JVC} distributions normalized to unity for positive (solid line) and negative (dashed line) truth b-jets. The irregular features reflect the binning used for the input JVC weight distributions.

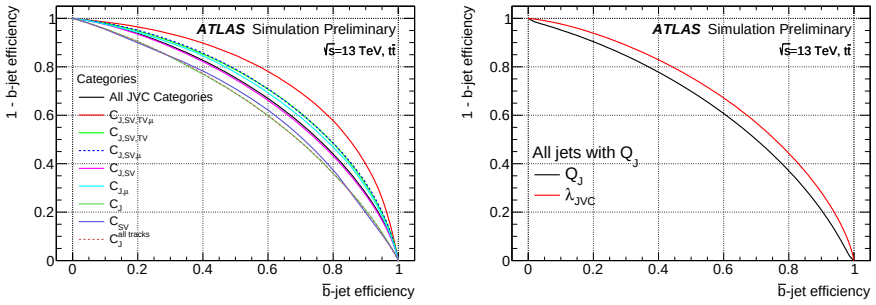
the performance, if a negative (positive) b-jet is identified to be the result of b-meson mixing, in the following it is considered as a positive (negative) b-jet.

Receiver Operating Characteristic (ROC) curves based on the final λ_{JVC} discriminant for all individual categories are compared to each other in Figure 4.8a. The overall improvement of the b-jet charge reconstruction is illustrated in Figure 4.8b, where the separation power of the final λ_{JVC} discriminant is compared to that of the Q_{J} variable: for a signal efficiency in the range 50–80%, the background rejection improves by 6–8%.

Figure 4.9 shows a more detailed break-down of the gain in the MVA-based categories due to the MLP discriminant compared to the basic charges used in the training.

Jet kinematics

The construction of the λ_{JVC} discriminant relies on quantities that have a substantial dependence on the jet kinematics. Variables such as mul-



(a) **(b)**

Figure 4.8: Comparison of background rejection curves for different categories (left) and a comparison of the gain in separation power of λ_{JVC} against the Q_{J} charge alone for all categories with a reconstructed Q_{J} (right).

tiplicities and momenta of tracks and muons produced within a jet are rather correlated to its transverse momentum. The track momentum then affects the track charge reconstruction efficiency, which drops significantly at high transverse momentum. On the other hand, jets with direction close to the edge of the angular acceptance of the Inner Detector are likely to lose some tracks due to the acceptance losses.

As a consequence, a dependence of the JVC performance on jet kinematics is expected and found, as can be seen in the plots in Figure 4.10, which show the λ_{JVC} separation power curves split in intervals of p_{T} and $|\eta|$ of the jet. The Jet Vertex Charge tagger performs better for lower jet p_{T} (up to 250 GeV), due the more efficient charge reconstruction of the tracks, whereas between 250 GeV and 500 GeV the separation power weakens significantly. On the other hand, the performance is rather stable up to $|\eta| < 2.1$, above which it slowly deteriorates as expected.

b-tagging working points

The Jet Vertex Charge algorithm exploits features of *b*-jets that are of great importance to *b*-tagging algorithms, hence an interplay between the JVC algorithm and *b*-tagging is expected.

The biggest effect of applying a *b*-tagging requirement on the jets is

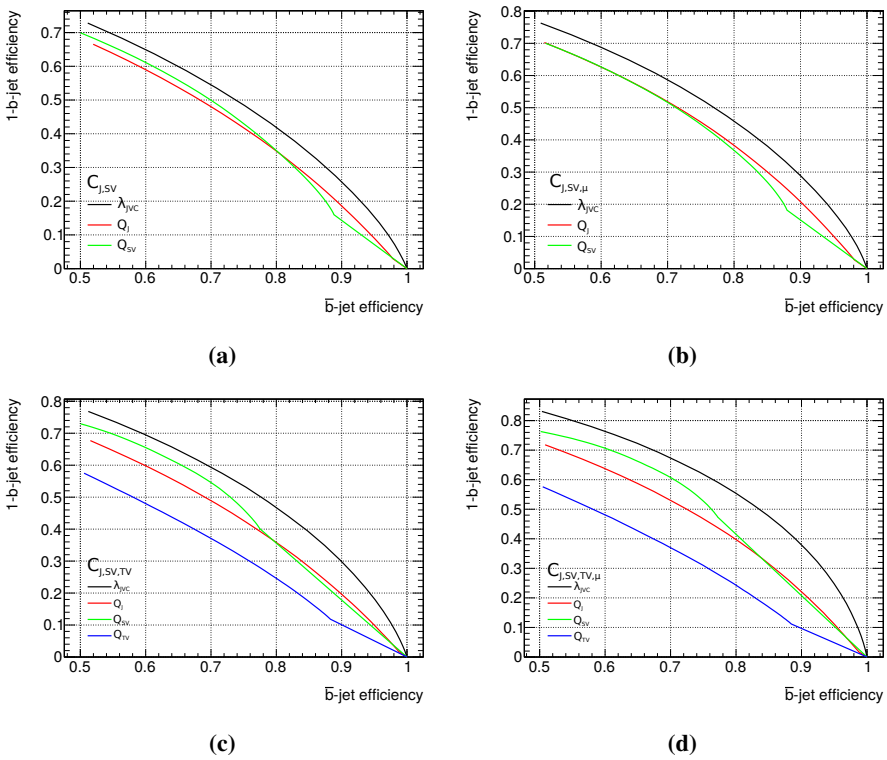
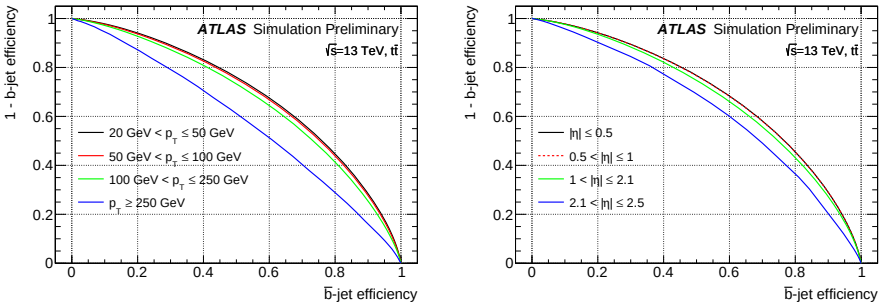


Figure 4.9: Background rejection curves based on λ_{JVC} compared to the available individual basic charges for the categories: $C_{\mathrm{J,\,SV}}$ (a), $C_{\mathrm{J,\,SV,}\,\mu}$ (b), $C_{\mathrm{J,\,SV,\,TV}}$ (c) and $C_{\mathrm{J,\,SV,\,TV,}\,\mu}$ (d).

the change in the relative fraction of the *b*-jet categories: the fraction of *b*-jets containing displaced vertices increases noticeably as the *b*-tagging requirement is tightened, as shown in Table 4.4.

The efficiencies of selecting the positively charged and rejecting the negatively charged truth *b*-jets for a range of representative requirements on the λ_{JVC} discriminant are presented in Table 4.5. For the loose requirement $\lambda_{\mathrm{JVC}} > 0$, both the efficiency and the rejection improve by 2%, with a higher improvement for higher JVC working points. The efficiency and rejection values are symmetric for negative λ_{JVC} cuts.



(a) **(b)**

Figure 4.10: Discrimination power in bins of the jet p_{T} (left) and jet $|\eta|$ (right).

Table 4.4: Relative abundance per category for various MV2c20 *b*-tagging efficiencies. Tightening the *b*-tagging efficiency changes the category composition, decreasing the fraction of jets with no displaced vertices.

Category	Untagged	85% eff.	77% eff.	70% eff.	60% eff.
C_{J}	11%	4.4%	2.8%	2.2%	1.6%
$C_{\mathrm{J},\,\mu}$	2.0%	0.95%	0.62%	0.49%	0.33%
C_{SV}	3.0%	3.3%	3.3%	3.2%	3.1%
$C_{\mathrm{J,\,SV}}$	53%	58%	58%	58%	57%
$C_{\mathrm{J,\,SV},\,\mu}$	10%	11%	11%	11%	11%
$C_{\mathrm{J,\,SV,\,TV}}$	15%	18%	19%	20%	22%
$C_{\mathrm{J,\,SV,\,TV},\,\mu}$	3.5%	4.1%	4.4%	4.6%	5.0%
C_{all}	1.3%	0.7%	0.5%	0.4%	0.3%

Table 4.5: Signal efficiencies and background rejection for representative working points on the λ_{JVC} discriminant obtained using the untagged truth b-jets, as well as those tagged at various working points of the MV2c20 b-tagging algorithm. The efficiency and rejection values are symmetric for negative λ_{JVC} cuts.

Tag configuration	$\lambda_{\text{JVC}} > 0.0$		$\lambda_{\text{JVC}} > 0.1$		$\lambda_{\text{JVC}} > 0.3$	
	$\bar{b}$ eff.	b rej.	$\bar{b}$ eff.	b rej.	$\bar{b}$ eff.	b rej.
Untagged	63.3%	63.8%	59.6%	68.2%	47.1%	77.8 %
MV2c20 at 85% eff.	63.9%	64.3%	59.4%	68.6%	48.1%	77.8 %
MV2c20 at 77% eff.	64.2%	64.5%	59.8%	68.8%	48.6%	77.9 %
MV2c20 at 70% eff.	64.5%	64.8%	60.1%	68.9%	49.1%	78.0 %
MV2c20 at 60% eff.	64.8%	65.2%	60.5%	69.2%	49.7%	78.1 %

4.2 Calibration analysis

In order to be able to use a tagger in an analysis, its performance needs to be evaluated both in real data and MC simulations, and potential differences between the two have to be taken into account.

This section describes the method used for the measurement of the performance of the JVC algorithm and the extraction of calibration scale factors (SF). The selected sample consists of $t\bar{t}$ candidate events with a single identified and isolated charged lepton (e or μ) in the final state and with exactly four jets, exactly two of which have to be b-tagged [125].

This choice is motivated by the fact that in such events sufficient kinematic constraints are present to allow for a pure reconstruction of the $t\bar{t}$ system. The charge of the lepton then provides a very clean reference to compare the reconstructed λ_{JVC} distribution of the b-jet of the leptonic top quark decay topology between data and simulated events. In this way it is also possible to extract the SF that can be used to correct the simulated λ_{JVC} distribution in analyses aiming to exploit this observable.

4.2.1 Data and simulated samples

The calibration analysis is based on data recorded by ATLAS in 2015 and 2016, for an integrated luminosity of 3.2 fb^{-1} and 32.9 fb^{-1} respectively, after requiring that all sub-detectors were functioning as expected and the LHC has declared that the conditions for stable beams were fulfilled.

All simulated events are processed through a software based on the GEANT4 toolkit [77] to simulate the response of the ATLAS detector and they are subsequently reconstructed using the same software as used for data. Samples generated with a fast simulation software, for which the calorimeter response is replaced by a parametrization of the shower shapes, are used to estimate modelling systematic uncertainties. In all the simulations, the top quark mass is fixed to the value of $m_t = 172.5$ GeV. The EvtGen (v1.2.0) program [132] is used to model the properties of bottom and charm hadron decays for all the non-SHERPA samples. Finally, to simulate the effects of pileup, additional interactions were generated using PYTHIA8 and overlaid on the simulated hard-scatter event. Each event is then reweighted to match the pileup profile seen in data.

Nominal $t\bar{t}$ + jets sample

The nominal $t\bar{t}$ + jets sample is produced using the POWHEG next-to-leading order (NLO) matrix element generator interfaced with the PYTHIA8 parton shower and hadronization processes (PS). It is normalized to the predicted theoretical cross section of $\sigma_{t\bar{t}} = 832^{+46}_{-51}$ pb, calculated with the Top++2.0 program [133] at the next-to-next-to-leading order (NNLO) in perturbative QCD, including resummation of next-to-next-to-leading logarithm (NNLL) soft gluon terms [134–138].

An important tune parameter of the POWHEG generator is h_{damp}, which controls the p_{T} of the first additional emission beyond the Born configuration. It is set to $1.5 \cdot m_{\text{t}}$, as it was found to provide the best description of data after the optimization process [139].

Alternative $t\bar{t}$ + jets samples

Alternative $t\bar{t}$ Monte Carlo samples are generated for checking the modelling of the $t\bar{t}$ system, as well as studying systematic uncertainties due to the matrix element generator, parton shower and hadronization model, or the initial and final state radiation.

The MC generator uncertainty for the hard process is evaluated by comparing the default POWHEG+PYTHIA8 sample to one generated by MADGRAPH5_aMC@NLO and interfaced to PYTHIA8. The parton shower and hadronization uncertainties are estimated by comparing the nominal POWHEG+PYTHIA8 sample to one where POWHEG is interfaced to Herwig7. Radiation systematics are evaluated by comparing the default POWHEG+PYTHIA8 sample with samples generated with "up" and "down" parameter variations.

The up variation has renormalization and factorization scales divided by two, the h_{damp} parameter up by a factor of two and shower radiation parameters up; on the contrary, the down variation has renormalization and factorization scales multiplied by two and the shower radiation parameters down. All variations are relative with respect to the nominal $t\bar{t}$ sample. Additional details can be found in Ref. [140].

Single top

The production of Wt, s-channel and the electroweak t-channel single top quark final states is modelled via the POWHEG generator. All the single top quark samples are generated at NLO accuracy and later normalized to the approximate NNLO theoretical cross sections [141–143]. The parton shower and hadronization process is modelled with the usage of PYTHIA6 [144].

Overlap between the $t\bar{t}$ and Wt final states is removed using the "diagram removal" procedure [145].

W/Z+jets and diboson

The W/Z+jets samples are generated using SHERPA. Matrix elements are calculated up to two partons at NLO and four partons at leading order (LO) using the Comix [146] and OPENLOOPS [147] matrix el-

ement generators and merged with the SHERPA parton shower [148]. These samples are later normalized to the NNLO cross-sections [149].

Samples of diboson events produced in association with jets are generated using SHERPA following the description presented in Ref. [150] and are normalized to their respective NLO cross-sections calculated by the generator.

Fake leptons

The only background not estimated using Monte Carlo simulation is the one composed of non-prompt and misidentified muons and electrons, collectively referred to as "fake" leptons. This background arises mostly from in-jet (semi-)leptonic decays of b- and c-hadrons and from photon conversions. The majority of this background is composed of multi-jet production with one fake lepton and non genuine missing energy. These processes are not well understood and therefore are difficult to model accurately, hence the estimation is done with a data-driven technique, the Matrix Method [151].

The calculations are based on the measurement of the number of events satisfying the nominal ("tight") lepton identification and isolation criteria as well as that satisfying more relaxed ("loose") criteria, together with measurements of the efficiencies for loose prompt and fake leptons to satisfy the tight criteria. In this analysis, the estimation is carried out separately for the 2015 and 2016 datasets; the loose criteria are defined by removing the isolation criteria and relaxing the identification criteria.

The efficiencies for loose leptons to pass the tight selection are measured in data for both real prompt and fake leptons. In this way it is possible to estimate the number of fake leptons passing the tight selection criteria by solving the system of equation:

$$\begin{aligned} N^l &= N^l_r + N^l_f \\ N^t &= \varepsilon_r N^l_r + \varepsilon_f N^l_f \end{aligned} \tag{4.3}$$

where N^l (N^t) is the number of events observed in data passing the loose (tight) lepton selection, N^l_r (N^t_f) is the number of events with a real (fake) lepton in the loose lepton sample, and ε_r (ε_f) is the fraction

of real (fake) leptons in the loose selection that also pass the tight one. For real prompt leptons the efficiency is measured in Z boson events, while for fakes it is estimated from events with low missing transverse momentum and low values of the reconstructed leptonic W boson transverse mass.

From a generalization of the formula to extract the number of fake leptons passing the tight selection, $N_f^t = \varepsilon_f N_f^l$, a weight is assigned to each of the selected events in the loose lepton data sample. The method thus provides a straightforward way to predict both the normalization and the kinematic distributions of this background.

4.2.2 Event selection and system reconstruction

Events are required to contain at least one reconstructed primary vertex candidate; if more than one vertex is found, the vertex with the highest sum of the squared transverse momenta of associated tracks is selected as the primary vertex.

Single-electron and single-muon trigger chains were used to collect events. The trigger chains are a combination in `OR` of triggers targeting low and high-p_{T} objects. The low-p_{T} triggers contain requirements on the lepton isolation and identification, whereas for the high-p_{T} triggers the criteria are relaxed, to recover efficiency due to the fact that energetic leptons tend to emit energy around them in the form of bremsstrahlung.

During the 2015 data taking period, the lowest p_{T} threshold was 24 GeV for the electron triggers and 20 GeV for the muon triggers. Both thresholds were raised during the 2016 campaign to 26 GeV. In particular, the higher p_{T} threshold during the 2016 data taking was determined by the increase in instantaneous luminosity between the two years.

An offline lepton p_{T} cut of 27 GeV, above the turn-on of the trigger, is applied and no additional lepton with $p_{\mathrm{T}} > 10\,\mathrm{GeV}$ must be present, in order to reject events in which the decay $W \to \tau \to \ell\nu_\ell$ has occurred. Every selected event must contain exactly four jets with $p_{\mathrm{T}} > 20$ GeV, exactly two of which must be b-tagged with the MV2c10 algorithm. Jets are b-tagged by requiring that the MV2c10 discriminant exceeds a fixed cut value, yielding a 70% efficiency for b-jets in simulated $t\bar{t}$ events and corresponding to rejection factors of 12 for c-jets and 380

for light-jets.

For each event, a final state reconstruction is performed with the Kinematic Likelihood Fitter (KLFitter) [152–154]. It is a reconstruction technique developed to reconstruct and determine the jet assignment for the $t\bar{t}$ decay products.

The jet-to-quark association is done by exploiting the decay topology of the $t\bar{t}$ system: in the single-lepton decay topology, the resulting tree level final state contains two b-quarks from the two top decays and two light or charm quarks from the W boson decay. A likelihood is used to properly assign these four jets to the true decay quarks; three out of the four jets selected in the final state are associated with the hadronic top decay, while the final fourth jet along with the charged lepton and neutrino are used to build the leptonic top. In the following, b_{lep} and b_{had} refer to the b-tagged jet associated with the leptonic and hadronic top quark decay.

The likelihood is built by multiplying Breit-Wigner terms (B) and transfer functions($\widetilde{W}$): the former are used to model the resonances present in the $t\bar{t}$ topology, i.e. the two W bosons and the two top quarks; while the latter are used to model the differences in the energy of the final state objects between the parton level and the reconstruction level. No b-tagging information is used in the evaluation of the likelihood, to avoid possible biases. The likelihood is therefore written as:

$$\begin{aligned} L \;= & B(m(q_1,q_2,b_{\text{had}})|m_{\text{t}},\Gamma_{\text{t}}) \cdot B(m(q_1,q_2)|m_W,\Gamma_W) \cdot \\ & B(m(\ell,\nu,b_{\text{lep}})|m_{\text{t}},\Gamma_{\text{t}}) \cdot B(m(\ell,\nu)|m_W,\Gamma_W) \cdot \\ & \widetilde{W}(E^{\text{meas}}_{\text{jet 1}}|E_{b_{\text{lep}}}) \cdot \widetilde{W}(E^{\text{meas}}_{\text{jet 2}}|E_{b_{\text{had}}}) \cdot \\ & \widetilde{W}(E^{\text{meas}}_{\text{jet 3}}|E_{q_1}) \cdot \widetilde{W}(E^{\text{meas}}_{\text{jet 4}}|E_{q_2}) \cdot \\ & \widetilde{W}^{\text{miss}}(E^{\text{miss}}_x|p_{x,\nu}) \cdot \widetilde{W}^{\text{miss}}(E^{\text{miss}}_x|p_{x,\nu}) \cdot \widetilde{W}^{\text{lep}}(p^{\text{meas}}_{\text{T}}|p^{\text{lep}}_{\text{T}}) \end{aligned} \tag{4.4}$$

Having exactly four jets in the final state means that a total of twelve permutations are considered, because the two jets from the hadronic W decay do not need to be distinguished.

On an event-by-event basis, the permutations are sorted based on the value of the likelihood to select the permutation that resembles the most the $t\bar{t}$ final state. The permutation with the highest likelihood, later

referred to as the best permutation, is adopted as the jet-to-parton assignment for the event.

Events with small values of the log-likelihood ($\mathrm{LLH}_{KLF} < -48$) are rejected. Finally, the two b-tagged jets must be associated with the b-quarks as determined by KLFitter.

Table 4.6 presents the KLFitter purity in bins of the b_{lep} p_{T}, estimated from the $t\bar{t}$ sample. The leptonic hemisphere of the $t\bar{t}$ decay offers a cleaner environment, due to less hadronic activity, which is reflected in the higher KLFitter purity of the b_{lep} throughout the whole p_{T} spectrum; for this reason only the b_{lep} is used in this analysis to measure the reconstructed λ_{JVC} distribution and derive calibration scale factors.

Table 4.6: KLFitter purity after selection for the five p_{T} bins of the analysis.

Matched topology	[20;30]	[30;60]	[60;90]	[90;140]	[140;200]
b_{lep}	84.36 %	85.97 %	88.60 %	94.72 %	99.15 %
b_{had}	81.61 %	81.12 %	83.78 %	90.35 %	95.83 %
both matched	78.82 %	80.18 %	83.35 %	90.14 %	95.68 %
b_{lep} and b_{had} swapped	10.76 %	12.07 %	10.30 %	4.83 %	0.62 %

In Table 4.7 the observed and expected event yields are shown, after the full event selection outlined above, as well as the truth flavour of the b_{lep}. Only statistical uncertainties are reported.

For the truth charge determination the hadron produced via the strong interaction is used, prior to any possible oscillation, in order to restore the correspondence between the charge of the lepton and the truth charge of the jet. The phenomenon of neutral b-meson oscillations is found to occur in approximately 15% of the cases.

From the data and MC yields in Table 4.7, obtained after the full event selection and including the system reconstruction, a clear overall normalization difference is apparent.

The procedure to calibrate the Jet Vertex Charge algorithm has been designed to be insensitive to normalization differences between data and simulation and for this reason a 2D reweighting procedure is applied. Given that the estimated non-$t\bar{t}$ component is quite small (less than 5% in the inclusive sample), it is assumed that the normalization difference

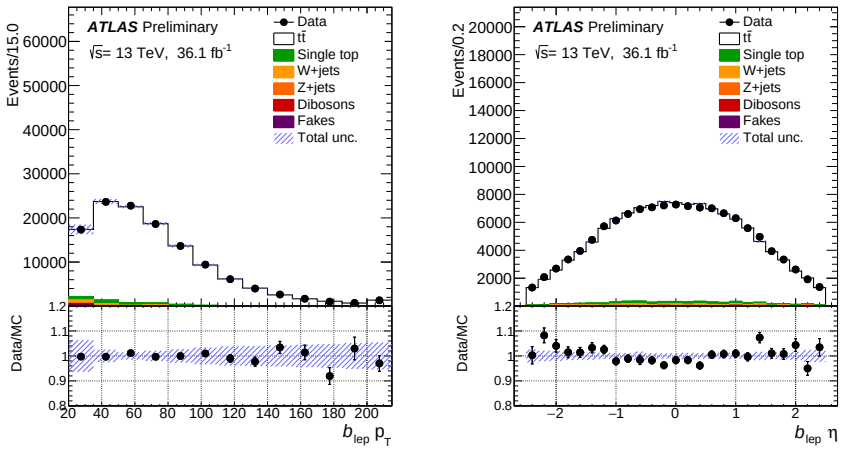
Table 4.7: Observed and expected yields in MC and data are shown after the full event selection, before and after the reweighting procedure explained in the text. The simulated flavour of the b_{lep} is shown as well. Only statistical uncertainties are shown.

Sample	Yields	After reweighting
$t\bar{t}$	122200 ± 210	116990 ± 200
single top	3384 ± 34	3384 ± 34
W+jets	1640 ± 120	1640 ± 120
Z+jets	487 ± 33	487 ± 33
dibosons	56 ± 4	56 ± 4
fakes	430 ± 120	430 ± 120
$\bar{b}$	62300 ± 150	59730 ± 140
b	63580 ± 150	60960 ± 150
$c/\bar{c}$	1260 ± 100	1240 ± 100
other	1050 ± 140	1050 ± 140
total MC	128200 ± 270	122990 ± 370
Data	122983	

in each p_{T} bin is due to the $t\bar{t}$ prediction, hence an additional weight is applied only on this sample. This weight is based on the values of the b_{lep} and b_{had} p_{T}, with the same p_{T} binning chosen to present the calibration results as a function of the p_{T} of the b_{lep}. The weights are derived in the same final phase space selected with the full selection criteria applied and range from 0.9 to 1.1 across the various p_{T} bins.

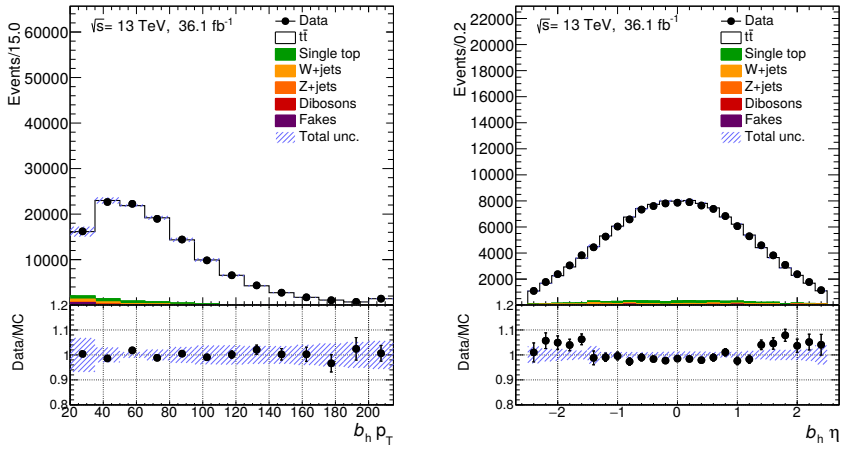
The p_{T} and η distributions after the reweighting procedure are shown in Figure 4.11 for the b_{lep} and in Figure 4.12 for the b_{had}. The uncertainty band consists of the sum in quadrature of the statistical and the systematic uncertainties, with the exclusion of the KLFitter uncertainty. The acceptance effects of the systematic uncertainties have been removed with the procedure described in Section 4.2.4.

Figure 4.13 shows the b_{lep} λ_{JVC} distribution after the reweighting. The reweighting has been found not to affect the shape of the λ_{JVC} distribution, giving confidence that it does not bias the final results of the analysis.



(a) **(b)**

Figure 4.11: Comparison between data and MC for the distributions of the b_{lep} p_{T} (left) and η (right) after the application of the reweighting. The uncertainty band consists of both the statistical and systematic uncertainties, with the systematic component computed with the method described in Section 4.2.4.



(a) **(b)**

Figure 4.12: Comparison between data and MC for the distributions of the b_{had} p_{T} (left) and η (right) after the application of the reweighting. The uncertainty band consists of both the statistical and systematic uncertainties, with the systematic component computed with the method described in Section 4.2.4.

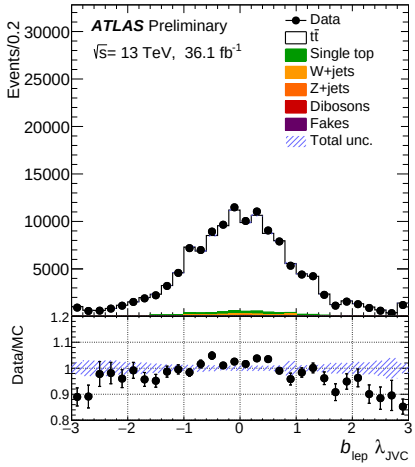


Figure 4.13: Comparison between data and MC for the λ_{JVC} distribution of the b_{lep}-jet after the reweighting is applied. The uncertainty band consists of both the statistical and systematic uncertainties, with the systematic component computed with the method described in Section 4.2.4.

4.2.3 Calibration strategy

In order to assess the charge of the b-quark from which the b-tagged jet originated, it is necessary to correlate it to the charge of the lepton; in fact, the lepton and the b-quark belonging to the top quark decaying leptonically will have a charge of opposite sign, whereas the charge of the b-quark belonging to the hadronically decaying top will have the same sign as that of the lepton. This correlation with the lepton charge is therefore used to further categorize the events into those having a b_{lep} with reconstructed positive or negative charge.

The kinematic reconstruction described above yields "raw" λ_{JVC} distributions of the b_{lep} candidates associated with negatively and positively charged leptons. The measurement of the fully corrected λ_{JVC} distribution for $\bar{b}$- and b-jets proceeds in a number of steps.

First, the charm and light-jet background contributions, as estimated from MC simulation, are subtracted from data events. After this step, the effects due to KLFitter misreconstruction are corrected with a procedure referred to as *unfolding*, which aims at removing only the KLFitter

impurity effects, not the detector resolution ones.

Labelling $g(q)$ the truth λ_{JVC} distribution for charge q, it is possible to express the reconstructed charge, $h(\pm)$, as:

$$\begin{cases} h(+) = \varepsilon\, g(\bar{b}) + (1-\varepsilon)\, g(b) \\ h(-) = (1-\varepsilon)\, g(\bar{b}) + \varepsilon\, g(b) \end{cases} \tag{4.5}$$

where ε is the KLFitter purity, which is identical for b- and $\bar{b}$-jets and is estimated from simulation, given that there is no straightforward method to measure it in data.

Solving this system removes the impurity effects due to KLFitter misreconstruction and provides access to the truth λ_{JVC} distributions for b- and $\bar{b}$-jets. Finally, the unfolded λ_{JVC} distribution for the negatively charged b_{lep} candidates $g(b)$, which is associated with b- rather than $\bar{b}$-jets, is *mirrored* to reflect the distribution for positively charged b_{lep} candidates. The mirroring procedure is nothing but a reflection of the λ_{JVC} distribution across a vertical line passing through the origin ($\lambda_{\text{JVC}} \to -\lambda_{\text{JVC}}$) and after the mirroring is performed, the distribution is added to the corresponding $g(\bar{b})$. It was verified that both the $g(b)$ and $g(\bar{b})$ are compatible with being each other's reflections.

The procedure described above is performed in five b_{lep} p_{T} bins, namely $\{20, 30, 60, 90, 140, 200\}$ GeV. Events for which the b_{lep} p_{T} is not in this range are not considered in the analysis.

4.2.4 Systematic uncertainties

Various sources of systematic uncertainties can affect the final results of the analysis and will be discussed in the following. These include detector related uncertainties as well as the modelling of the physical processes.

The effect of each uncertainty is evaluated by recomputing new SF by replacing the nominal Monte Carlo sample with a modified sample affected by a single systematic uncertainty variation of $\pm 1\sigma$ for the up/down variation respectively. Uncertainties affect both the acceptance and the shape of the samples. On the other hand, the $t\bar{t}$ signal's cross section uncertainty is not relevant for this analysis, given that the signal is reweighted to reproduce the data, as described in Section 4.2.2. More

generally, acceptance effects are removed by normalizing the modified templates so that their yields will match the total yield observed in data. Then, for each systematic source, the envelope of the new SF around the nominal SF is taken as the uncertainty associated with the systematic source under consideration.

All the systematic uncertainties are then summed in quadrature to obtain the final uncertainty on the result.

KLFitter uncertainty

The calibration analysis relies heavily on the matching done by KLFitter; therefore, an assessment of the impact of its impurity is crucial.

The KLFitter purity is taken from simulation and, in principle, it can differ in data. Given the impossibility to estimate it in data and in order to cover possible differences between KLFitter purity in data and MC, a specific systematic is introduced. It is derived by comparing the SF obtained with the nominal selection of the analysis and a tightened selection intended to further reduce the impurity, hence obtaining an estimate of how much the result will change in different conditions.

The discriminating variable chosen for this purpose is the mass of the hadronic top candidate, reconstructed by swapping the assignments of the two b-jets. The distribution is shown in Figure 4.14. A peak is clearly visible at the position of the top mass for the wrongly matched b-jets, which indicates that in these cases swapping the two b-jets restores the correct parton-to-jet matching. Furthermore, the visible dip in the same mass range is a consequence of the KLFitter reconstruction: for events that are correctly matched by KLFitter, in the swapped assignment the new hadronic top quark mass is less likely to agree with the top quark mass. By rejecting events in the mass window between 120 GeV and 220 GeV, the KLFitter matching purity increases in all the p_{T} bins, as can be seen in Table 4.8.

In Figure 4.15 the double ratio of data and MC for the two selections is shown. Only the statistical uncertainty is shown and the correlation between the two subsamples it taken into account in its computation. The two selections provide compatible SF in most (but not all) of the bins, therefore the deviation of their double ratio from unity is taken as the relative uncertainty associated with the KLFitter impurity.

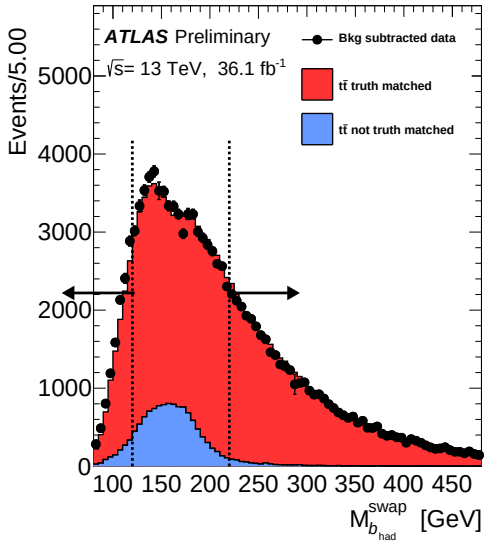


Figure 4.14: Mass of the top quark decaying hadronically, reconstructed by swapping the b_{lep} and the b_{had}. The red histogram ("$t\bar{t}$ truth matched") represents events for which the identification of the b_{lep} by KLFitter correctly matches the corresponding simulated b_{lep} jet, whereas the blue histogram ("$t\bar{t}$ non truth matched") collects all the remaining events. A peak at the value of the top quark mass is visible for the wrongly matched events. Events falling in the mass range identified by the dashed vertical lines at 120 GeV and 220 GeV are rejected in order to estimate the systematic uncertainty associated with KLFitter.

$t\bar{t}$ modelling

The $t\bar{t}$ sample is the most important one for the analysis, hence dedicated systematics are employed in order to cover possible mismodelling. These uncertainties are estimated by replacing the nominal MC sample with the alternative ones described in Section 4.2.1.

As already outlined previously, in order to assess the uncertainty due to the Monte Carlo generator choice for the simulation of the hard scatter, the POWHEG+PYTHIA8 sample is compared to the sample generated with MADGRAPH5_aMC@NLO interfaced with PYTHIA8.

Table 4.8: KLFitter purity for the five p_{T} bins of the analysis after the nominal selection and the additional cut on the mass of the hadronic top candidate with the swapped b-jet assignments.

Matched topology	[20;30]	[30;60]	[60;90]	[90;140]	[140;200]
b_{lep}	88.75 %	93.69 %	97.38 %	98.65 %	99.51 %
b_{had}	84.76 %	87.70 %	93.24 %	95.37 %	96.51 %
both matched	81.64 %	86.53 %	92.83 %	95.17 %	96.37 %
b_{lep} and b_{had} swapped	5.89 %	4.19 %	1.84 %	1.01 %	0.28 %

The uncertainty related to the choice of parton shower and hadronization model is derived by comparing the POWHEG+PYTHIA8 nominal sample to the prediction obtained by using the same matrix element generator, but interfaced with a different parton shower, Herwig7. These two PS generators use different algorithms for the parton shower and for the hadronization modelling, therefore by comparing the two samples a systematic uncertainty on the fragmentation model is included as well. In particular, Herwig7 uses an angular ordering as the evolution variable of the shower process, while PYTHIA8 ordering is done based on the $p_{\perp}$ variable, the component of the momentum of the emitted parton perpendicular to the momentum of the incoming initial parton.

An uncertainty associated with the modelling of the initial and final state radiation is assessed by the usage of two dedicated "up" and "down" samples generated with POWHEG+PYTHIA8 samples.

Other background modelling systematic

Given the small contribution of non-$t\bar{t}$ samples in the final event yields, a generic, conservative 50% normalization uncertainty is assigned to each of the different Monte Carlo samples, as well as to the data-driven fake lepton estimate.

Experimental systematic uncertainties

The uncertainties related to the reconstructed objects have been described in Chapter 3.

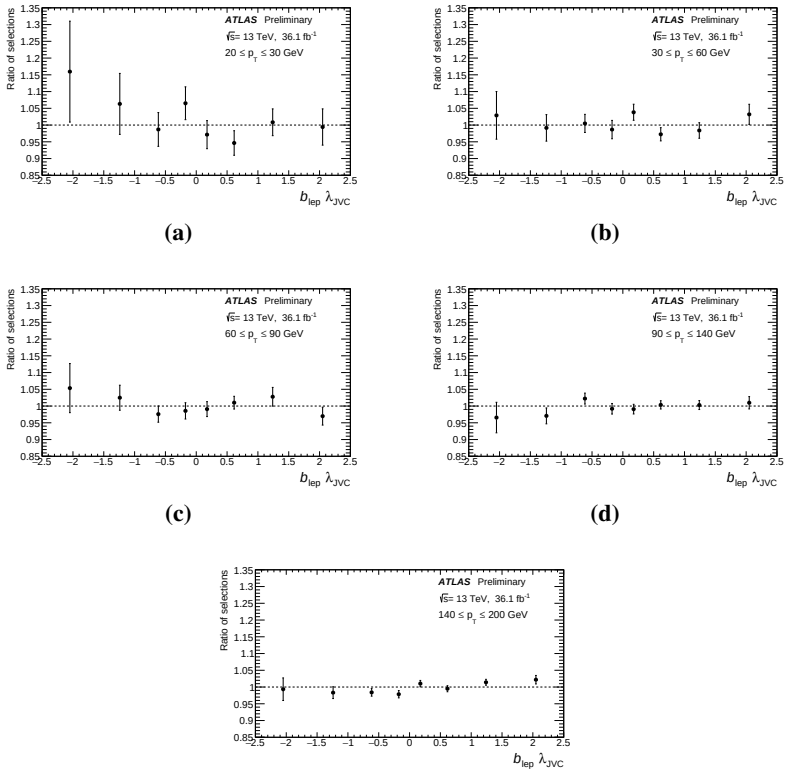


Figure 4.15: Double ratio of the SF obtained with the two selections for the five p_{T} bins of the analysis. Only the statistical error on the double ratio is shown and the correlation between the two subsamples is taken properly into account.

The uncertainty in the combined 2015+2016 integrated luminosity is 2.1%. It is derived following a methodology similar to that detailed in Ref. [155], from a calibration of the luminosity scale using x–y beam-separation scans performed in August 2015 and May 2016.

A variation in the pileup reweighting of MC events is included to cover the uncertainty in the ratio of the predicted and measured inelastic cross-sections in the fiducial volume defined by $\mathrm{M}_X > 13$ GeV, where

M_X is the mass of the hadronic system [156].

Uncertainties associated with jets arise from the efficiency of the pileup-jet rejection based on the JVT variable, as well as the jet energy resolution (JER) and jet energy scale (JES).

JES and its uncertainty were derived by combining information from test-beam data, LHC collision data and simulation, as outlined in Section 3.3.

Flavour tagging efficiencies in simulated samples are corrected in order to match efficiencies measured in data, thus uncertainties related to the calibration of these corrections are considered as well. Correction scale factors are derived for jets originating from gluons and light, c- and b-quarks separately in dedicated calibration analyses.

These uncertainties are then used as input into an eigenvector variation model with a reduction scheme such that only substantial variations are treated separately while all small ones are combined together[4]. A total of 6, 4 and 16 independent eigenvectors are considered for b, c and light-jets respectively. An additional uncertainty is considered for the extrapolation of the flavour tagging SFs for jets with p_{T} outside the kinematic range used for their measurement. Lastly, jets from hadronic τ lepton decays are considered as c-jets for the mis-tag rate corrections and systematic uncertainties. An additional source of systematic uncertainty is considered on the extrapolation from c-jets to these τ-jets

Scale factors are used to correct differences between data and simulation for the lepton identification, isolation, trigger and reconstruction, as well as their momentum scale and resolution, as outlined in Section 3.2. Their uncertainties are considered as a source of systematic uncertainty as well.

Finally, three genuine sources of systematic uncertainties are considered for the missing energy, associated with the calculation of the scale and resolution of the soft term, as described briefly in Section 3.4.

[4] A detailed description of the method can be found in Appendix B of Ref. [157].

Breakdown of systematic uncertainties

Tables 4.9 to 4.13 show the impact of systematic uncertainties, grouped based on their source, on the total error of the SF, per p_{T} bin in each of the JVC bins. The values are the absolute, not relative, difference with respect to the central value of the SF. Only rows for which at least one of the bin presents a deviation larger than 0.01 are explicitly reported, but all contributions are included in the last row of the tables, showing the total systematic uncertainty.

The largest deviations from the nominal value of the SF come from the KLFitter uncertainty, especially in the lowest λ_{JVC} bin, and the $t\bar{t}$ modelling systematics.

The flavour tagging systematic uncertainties, as well as uncertainties related to the lepton do not have a big impact in all the λ_{JVC} and p_{T} bins, whereas a bigger, even though not substantial impact comes from some of the JER/JES in a few λ_{JVC} bins.

Table 4.9: Variation due to the systematic uncertainties on the final SF per source of uncertainty, per λ_{JVC} bin, in the first p_{T} bin ($20 < p_{\mathrm{T}} < 30$ GeV) of the analysis.

Syst. name	$[-2.5,-1.6]$	$[-1.6,-0.88]$	$[-0.88,-0.35]$	$[-0.35,0]$	$[0,0.35]$	$[0.35,0.88]$	$[0.88,1.6]$	$[1.6,2.5]$
KLFitter	0.16	0.06	0.01	0.07	0.03	0.05	0.01	0.01
Jet unc.	0.05	0.06	0.07	0.03	0.01	0.03	0.02	0.04
Flav. tag.	0.02	0.02	0.07	0.02	0.01	0.02	0.01	0.02
Elec. unc.	0.00	0.00	0.03	0.01	0.00	0.01	0.01	0.01
Muon unc.	0.01	0.00	0.05	0.01	0.00	0.02	0.01	0.01
MET	0.03	0.01	0.03	0.01	0.00	0.01	0.01	0.01
PileUp	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.01
Fake leptons	0.05	0.04	0.01	0.00	0.02	0.00	0.01	0.02
$t\bar{t}$ model	0.11	0.09	0.04	0.02	0.02	0.02	0.03	0.04
W+jets XS	0.03	0.03	0.01	0.01	0.00	0.00	0.01	0.02
Z+jets XS	0.02	0.00	0.01	0.00	0.00	0.00	0.00	0.01
Stat. unc.	0.12	0.08	0.05	0.05	0.04	0.04	0.04	0.05
Tot syst unc.	0.21	0.14	0.13	0.08	0.04	0.07	0.05	0.07

Table 4.10: Variation due to the systematic uncertainties on the final SF per source of uncertainty, per λ_{JVC} bin, in the second p_{T} bin ($30 < p_{\text{T}} < 60$ GeV) of the analysis.

Syst. name	$[-2.5,-1.6]$	$[-1.6,-0.88]$	$[-0.88,-0.35]$	$[-0.35,0]$	$[0,0.35]$	$[0.35,0.88]$	$[0.88,1.6]$	$[1.6,2.5]$
KLFitter	0.03	0.01	0.00	0.01	0.04	0.03	0.02	0.03
Jet unc.	0.04	0.01	0.01	0.01	0.00	0.01	0.00	0.01
Elec. unc.	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Muon unc.	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$t\bar{t}$ model	0.13	0.06	0.04	0.02	0.01	0.02	0.03	0.06
Stat. unc.	0.06	0.03	0.02	0.02	0.02	0.02	0.02	0.02
Tot syst unc.	0.15	0.07	0.04	0.03	0.04	0.03	0.03	0.07

Table 4.11: Variation due to the systematic uncertainties on the final SF per source of uncertainty, per λ_{JVC} bin, in the third p_{T} bin ($60 < p_{\text{T}} < 90$ GeV) of the analysis.

Syst. name	$[-2.5,-1.6]$	$[-1.6,-0.88]$	$[-0.88,-0.35]$	$[-0.35,0]$	$[0,0.35]$	$[0.35,0.88]$	$[0.88,1.6]$	$[1.6,2.5]$
KLFitter	0.05	0.02	0.02	0.01	0.01	0.01	0.03	0.03
Jet unc.	0.06	0.03	0.02	0.01	0.00	0.01	0.02	0.01
$t\bar{t}$ model	0.16	0.07	0.04	0.02	0.00	0.02	0.03	0.06
Stat. unc.	0.07	0.03	0.02	0.02	0.02	0.02	0.02	0.02
Tot syst unc.	0.18	0.08	0.05	0.03	0.01	0.02	0.05	0.07

Table 4.12: Variation due to the systematic uncertainties on the final SF per source of uncertainty, per λ_{JVC} bin, in the fourth p_{T} bin ($90 < p_{\text{T}} < 140$ GeV) of the analysis.

Syst. name	$[-2.5,-1.6]$	$[-1.6,-0.88]$	$[-0.88,-0.35]$	$[-0.35,0]$	$[0,0.35]$	$[0.35,0.88]$	$[0.88,1.6]$	$[1.6,2.5]$
KLFitter	0.03	0.03	0.02	0.01	0.01	0.00	0.00	0.01
Jet unc.	0.03	0.01	0.01	0.01	0.01	0.00	0.00	0.01
Flav. tag.	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$t\bar{t}$ model	0.08	0.04	0.02	0.02	0.00	0.01	0.02	0.03
Stat. unc.	0.07	0.04	0.02	0.02	0.02	0.02	0.02	0.02
Tot syst unc.	0.10	0.06	0.03	0.02	0.01	0.01	0.02	0.03

Table 4.13: Variation due to the systematic uncertainties on the final SF per source of uncertainty, per λ_{JVC} bin, in the fifth p_{T} bin ($140 < p_{\text{T}} < 200$ GeV) of the analysis.

Syst. name	$[-2.5,-1.6]$	$[-1.6,-0.88]$	$[-0.88,-0.35]$	$[-0.35,0]$	$[0,0.35]$	$[0.35,0.88]$	$[0.88,1.6]$	$[1.6,2.5]$
KLFitter	0.01	0.02	0.02	0.02	0.01	0.01	0.01	0.02
Jet unc.	0.04	0.04	0.04	0.04	0.04	0.03	0.04	0.04
Flav. tag.	0.03	0.04	0.04	0.04	0.04	0.03	0.04	0.03
Elec. unc.	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01
Muon unc.	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.02
MET	0.02	0.01	0.02	0.02	0.02	0.01	0.01	0.01
Fake leptons	0.02	0.01	0.01	0.02	0.00	0.00	0.01	0.01
$t\bar{t}$ model	0.12	0.09	0.07	0.05	0.03	0.02	0.07	0.12
Stat. unc.	0.12	0.06	0.04	0.04	0.04	0.03	0.04	0.05
Tot syst unc.	0.14	0.12	0.10	0.09	0.07	0.06	0.10	0.13

4.2.5 Calibration results

This section presents the measurement of the Jet Vertex Charge distributions in data and the corresponding data-to-simulation scale factors.

Figure 4.16 shows the b_{lep} λ_{JVC} distributions in the five p_{T} bins of the analysis, as well as the contributions of the simulated distributions by the different truth jet flavours.

In dark blue is presented the distribution for truth $\bar{b}$-jets identified as having a positive charge by the jet-lepton correlation ("$\bar{b}$ pos" in the legend), whereas the light blue distribution represents the events for which the truth $\bar{b}$-jet is identified as having a negative charge ("$\bar{b}$ neg"). The dark (light) red histogram represents the λ_{JVC} distribution for truth b-jets identified as having a negative (positive) charge, "b neg" ("b pos") in the legend. The contribution of c-jets (yellow) and light-jets (green) is negligible in all but the first b_{lep} p_{T} bin. The uncertainty band corresponds to the sum in quadrature of the statistical and systematic uncertainties, with the exception of the KLFitter systematic. The systematic uncertainty band comprises only shape effects, with acceptance effects removed by using the procedure described in Section 4.2.4.

The effect of the correction procedure described in Section 4.2.3 can be seen in Figure 4.17, where the final λ_{JVC} distributions are shown after the unfolding correction steps are applied. The data/MC ratio in the bottom panel shows a good, but not perfect modelling of data offered by the simulation. Therefore this ratio is considered as a calibration

scale factor, which can be used to correct λ_{JVC} distributions in simulated samples in physics analyses exploiting this observable. In order to display more clearly the SF, the bottom panel is also presented separately in Figure 4.18. The systematic uncertainty band in both figures comprises only shape effects, but includes the uncertainty associated with KLFitter.

Table 4.14 presents the algorithm performance estimated from the simulated $t\bar{t}$ sample. It presents the efficiency of correctly tagging the charge of the b_{lep} based on the charge of the lepton. The efficiencies are measured inclusively and for the five p_{T} bins in the same phase space selected for the calibration analysis, with the additional requirement that the b_{lep} is correctly matched at truth level with the b-jet from the leptonic top quark. As a representative cut value, jets with $\lambda_{\mathrm{JVC}} > 0$ are tagged as having a positive charge.

The effect of oscillations is visible in the last two rows of the table. In this case, the tagged charge of the b_{lep} is compared to the truth charge of the b-meson before and after the oscillations occur. Given that the algorithm is trained on the weakly decaying b-hadron, higher efficiencies are expected after the mixing occurred.

Table 4.14: Jet Vertex Charge performance estimated from the simulated $t\bar{t}$ sample, based on the charge of the lepton. The selection is the same as for the calibration analysis, with the additional requirement that the b_{lep} is correctly matched at truth level with the b-jet from the leptonic top quark. Jets are tagged as having a positive charge if $\lambda_{\mathrm{JVC}} > 0$. "OS" stands for "opposite sign".

	[20;30]	[30;60]	[60;90]	[90;140]	[140;200]	Inclusive
b_{lep} OS lepton	64.10 %	64.65 %	65.44 %	64.90 %	64.77 %	64.91 %
b_{lep} OS b_{had}	54.31 %	54.14 %	54.82 %	54.62 %	54.87 %	54.51 %
before mixing	64.40 %	65.20 %	65.74 %	65.16 %	64.68 %	65.26 %
after mixing	68.04 %	68.56 %	68.58 %	67.98 %	66.94 %	68.31 %

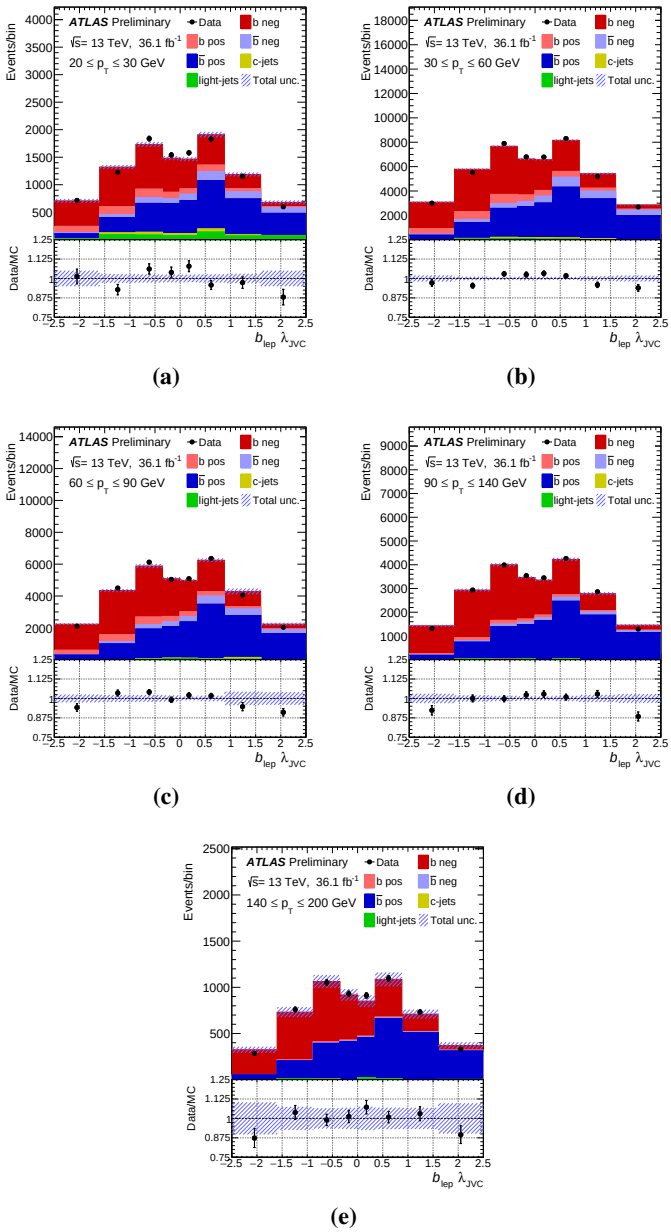


Figure 4.16: λ_{JVC} distribution, before corrections, split into the different flavour components for the 5 p_{T} bins of the analysis. The uncertainty band corresponds to the statistical $\oplus$ systematic uncertainties, with the latter comprising only shape effects.

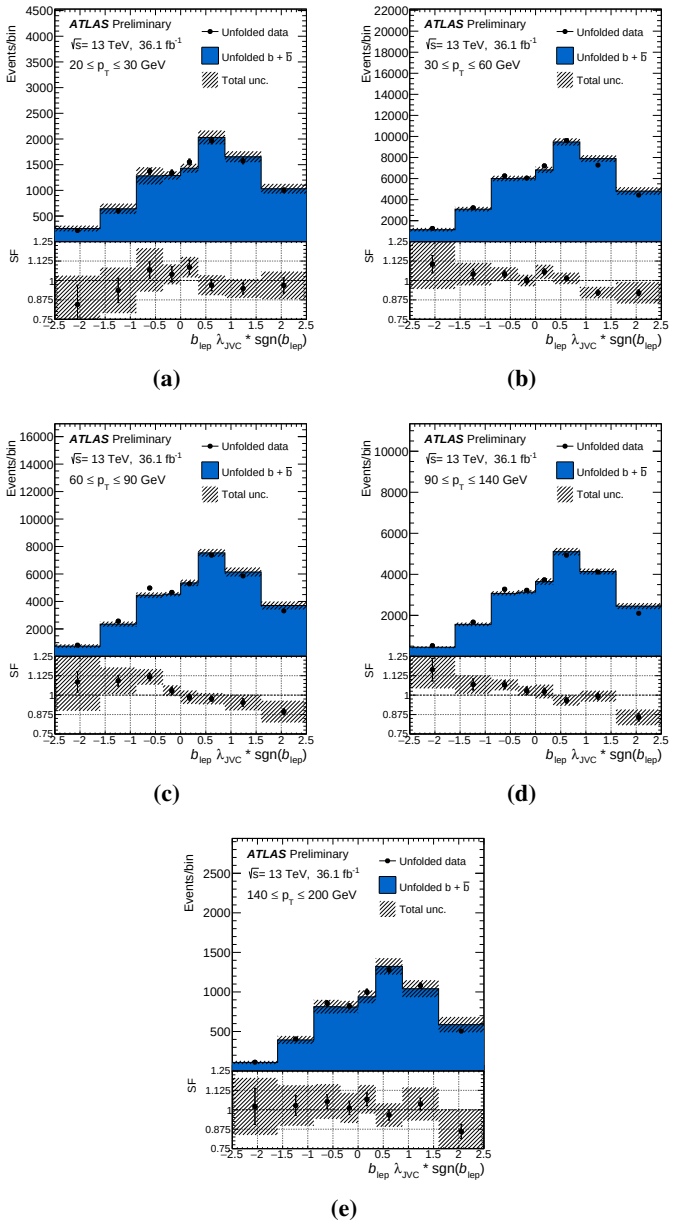
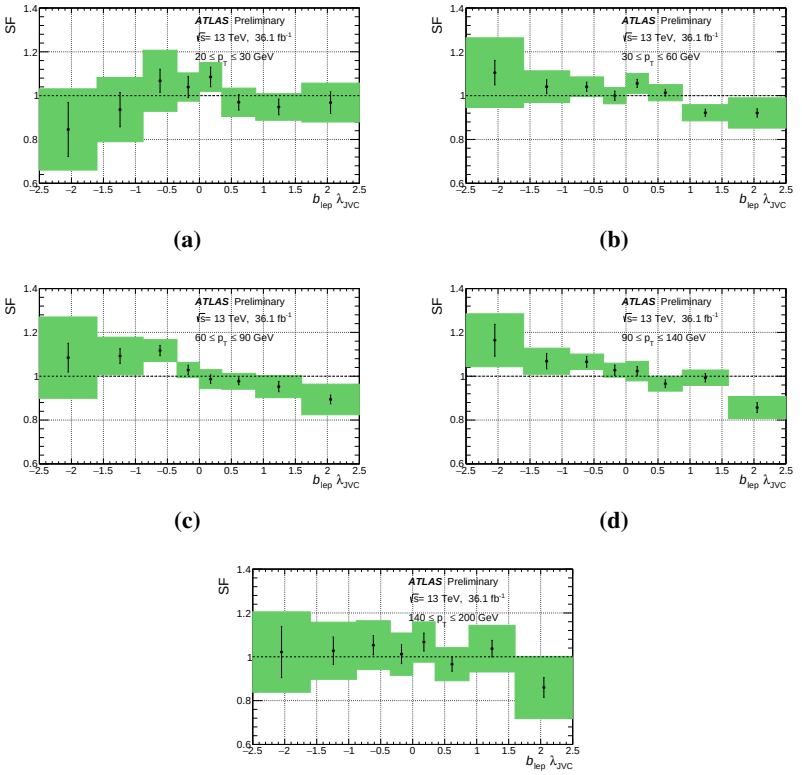


Figure 4.17: Final λ_{JVC} distribution for the b_{lep} in the 5 p_{T} bins of the analysis. The distributions shown correspond to the sum of the unfolded $g(\bar{b})$ and unfolded $g(b)$. The multiplication by sgn(b_{lep}) indicates that the $g(b)$ distribution has been reflected about 0 for both data and simulation. The uncertainty band corresponds to the statistical $\oplus$ systematic uncertainties, with the latter comprising only shape effects.



(a) **(b)**

(c) **(d)**

(e)

Figure 4.18: SF for the λ_{JVC} distribution split into the different flavour components for the 5 p_{T} bins of the analysis. The uncertainty band corresponds to the statistical $\oplus$ systematic uncertainties, with the latter comprising only shape effects.

The road to $t\bar{t}H$

5

In the Standard Model, the Higgs boson is the particle associated with the field that is responsible for particles to acquire mass via the Brout-Englert-Higgs mechanism.

After its discovery in 2012 by the ATLAS and CMS Collaborations and the first measurements of its interactions, which permit to probe the mechanism of spontaneous symmetry breaking, it is still of fundamental importance to determine precisely of all of its properties, as well as to observe the missing production and decay modes in order to probe the internal consistency of the SM or find deviations from its predictions.

As discussed in the end of the first chapter, among all the properties the top Yukawa coupling plays a special role and the $t\bar{t}H$ channel is the best candidate to have a direct access, at tree level, to this coupling. On the contrary, other production modes, such as the ggF or decay modes involving loops, offer just an indirect way of measuring it, given that all particles, even particles associated with physics beyond the SM, can enter such loops.

The Standard Model $t\bar{t}H$ production cross section is equal to 0.5 pb, which is roughly 1% of the total Higgs boson production cross section. The tiny fraction of the overall cross section pushes to look for it by exploiting as many Higgs decay modes as possible. The decay mode with the highest BR is into a pair of bottom quarks, $H \to b\bar{b}$ (58%), which incidentally also contributes to the measurement of the bottom Yukawa coupling. Other decay modes exploited are the Higgs boson decaying into a pair of WW, ZZ bosons and $\tau\tau$ leptons, as well as the Higgs decay into a pair of photons.

In this chapter, the search for $t\bar{t}H$ will be presented, with the main focus on the $H \to b\bar{b}$ decay. A general introduction to the $t\bar{t}H(b\bar{b})$ anal-

ysis strategy will be given, first describing a preliminary version of the analysis, which was performed only on a subset of the total recorded data, corresponding to 13.2 fb^{-1}. This first version of the analysis was presented at the ICHEP conference in Chicago in 2016, in which the possibility to improve its sensitivity with the usage of Jet Vertex Charge tagger was studied, and will be presented in Section 5.3.

The latest result, with the full dataset collected by the ATLAS experiment in 2015 and 2016, referred to as the "paper analysis", will be discussed in Section 5.6. Finally, the combination with the other channels and the evidence for the Higgs boson production in association with top quarks will be presented in Section 5.7.

5.1 General discussion about analysis strategy

The presence of a pair of top quarks in the final state offer a distinctive signature for the identification of the $t\bar{t}H$ process. The top quark is not only the heaviest of all the quarks in the SM, but its lifetime is much shorter than the hadronization time ($\tau = 1/\Gamma_{top} < 1/\Lambda_{\text{QCD}}$), so that properties of a *free* quark can be measured directly without the extra complications coming from non-perturbative effects due to the hadronization.

The top quark decays almost exclusively into a b-jet, as the CKM element $|V_{tb}| \sim 1$, and a W boson; for this reason the decay mode of the on-shell W boson is used to identify the different topologies of final state: dileptonic (DIL), semileptonic or single-lepton (SL) and fully hadronic final states. In the first case, both W bosons decay into a lepton and the corresponding neutrino, $W^{\pm} \to l^{\pm}\nu_l$, in the semileptonic mode only one of the W bosons decays leptonically and the second one decays into a quark-antiquark pair and finally the full hadronic mode occurs when both W bosons decay into a quark-antiquark pair.

The hadronic decay of the W boson happens in approximately $2/3$ of the cases, with the quarks being often lighter than a b-quark, as $|V_{cb}|$ is small and $|V_{ub}|$ is even smaller, such that the jets produced are often light and c-jets. In the other $1/3$ of the cases, it decays into a lepton-neutrino pair, with the three lepton flavours having the same probability, due to lepton universality.

In Figure 5.1 the BR for the different topologies of a top quark pair decay can be seen. The biggest BR is in the full-hadronic mode, whereas the dileptonic mode has the smallest, but also the cleanest signature and the semileptonic topology sits in-between the two cases.

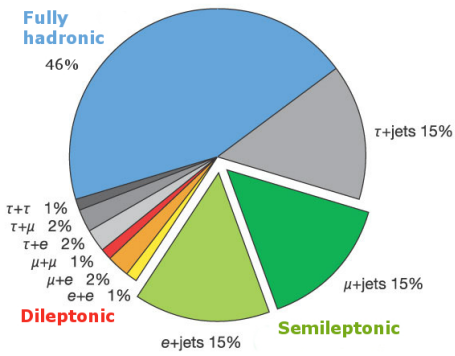


Figure 5.1: Branching ratios of a top quark pair decay.

In the following, unless specified otherwise, only decays into light leptons (e or μ) will be considered as a leptonic W boson decay; decays into τ leptons will be not considered, unless the τ subsequently decays into (light) leptons.

The $t\bar{t}H(b\bar{b})$ analysis targets only events falling into the SL and DIL channels.

One of the main experimental challenges rises from the difficulty to reconstruct and identify all the objects in the final state; this is due to both the combinatorial ambiguity that makes it difficult to efficiently reconstruct the Higgs boson mass and due to the presence of one (two) neutrino(s) in the SL (DIL) final state.

Additional experimental challenges come from the correct description of the dominant background of the $t\bar{t}$ + jets, in particular when the extra jets originate from a heavy flavour decay, e.g. the final state $t\bar{t} + \geq 1b$ or its subcomponent $t\bar{t} + b\bar{b}$. Feynman diagrams for the $t\bar{t}H(b\bar{b})$ and $t\bar{t} + b\bar{b}$ processes are shown in Figure 5.2.

The $t\bar{t}$ + jets background poses a serious challenge, as its inclusive cross section is roughly 1600 times larger than the $t\bar{t}H$ cross section; moreover events with additional heavy-flavour jets dominate the background composition in the signal regions.

More importantly, the large uncertainty in the MC simulations represents one of the main bottlenecks of the searches for the $t\bar{t}H(b\bar{b})$ final state, therefore state-of-the art theory predictions for the production of

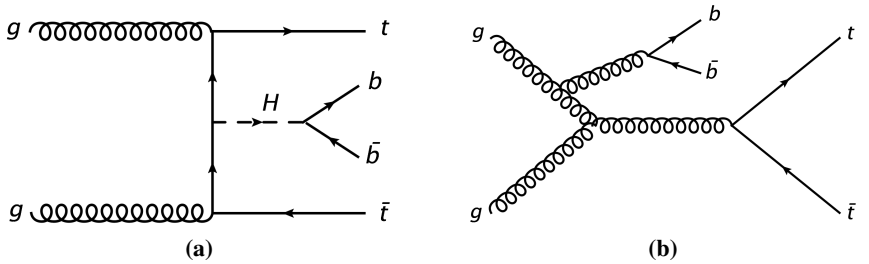


Figure 5.2: Tree level Feynman diagrams for the production of the Higgs boson in association with a top quark pair ($t\bar{t}H$) and the subsequent decay of the Higgs to a $b\bar{b}$ (left) and for the main background $t\bar{t}+b\bar{b}$ (right).

$t\bar{t}+jj$ and its subcomponents $t\bar{t}+c\bar{c}$ and $t\bar{t}+b\bar{b}$ are a key ingredient to the final sensitivity of the analysis. This is particularly true for the irreducible $t\bar{t}+b\bar{b}$ background, where theory predictions play an especially important role. NLO calculations for $t\bar{t}+b\bar{b}$ [158, 159] and $t\bar{t}+jj$ [160] production can heavily reduce perturbative uncertainties from 70–80% down to 15–20% [161].

One of the theoretical challenges in the modelling of the $t\bar{t}+b\bar{b}$ process comes from the fact that b-quarks are massive, while gluon and light quarks can be safely considered massless in the calculations. This affects various properties of the process, because the gluon splitting contributions for the $t\bar{t}+b\bar{b}$ dominate over the double initial state radiation for the $t\bar{t}+jj$ in the final phase space selected by the analysis At the same time, the description of the gluon splitting into massive quarks is a critical component, in particular for low angular separations, in which fixed order calculation have similar precision to the analytical parton shower programs. Nevertheless, measurements of the $t\bar{t}+b$ and $t\bar{t}+b\bar{b}$ fiducial cross sections, as well as $t\bar{t}+b\bar{b}/t\bar{t}+jj$ cross section ratio have been performed by ATLAS and CMS [162, 163] to aid the theoretical predictions for the $t\bar{t}$ +b process [164].

After an inclusive selection, events are then categorized according to their jet and b-jet multiplicity in different signal (SR) and control regions (CR). The signal regions are identified by their expected $S/\sqrt{B}$ and S/B ratios; high values of both figures of merit flag the region as

signal region[1].

In the signal regions, several techniques are employed in order to reconstruct the final state and identify the origin of the various object, i.e. identify the two b-tagged jets coming from the $H \to b\bar{b}$ decay as well as a full reconstruction of the top pair system. After the event reconstruction, multivariate output discriminants are used to further classify events into more or less signal-like. This second layer uses Boosted Decision Trees (BDT) and it is known as "classification BDT". The advantages of having both signal and background regions lies in the fact that, while different signal fractions in the SR can maximize the sensitivity of the statistical combination, the CR are used to improve the knowledge of the systematic uncertainties and normalization of the various background components, directly, in events with a topology close to the signal, hence reducing their impact.

This is achieved by using a profile likelihood fit, with all the analysis regions used as input. A more detailed discussion of the method will be exposed in Section 5.5.

5.2 Signal and background modelling

The simulated samples used in the Jet Vertex Charge calibration analysis described in Chapter 4 are used in this analysis as well; for this reason, only additional samples and different treatments will be described in the following. For further details, "my twenty-five readers"[2] are referred to Section 4.2.

All simulated events are processed through the full simulation of the ATLAS detector based on GEANT4. Samples generated with a fast simulation software, for which the calorimeter response is replaced by a parametrization of the shower shapes, are used to estimate modelling systematic uncertainties. Simulated events are subsequently reconstructed using the same software also used for data.

[1] Both figures of merit are useful in this context, as the usual significance expression, $S/\sqrt{B}$, is only valid under the assumption that systematic uncertainties are small compared to the statistical ones, which is not necessarily the case for regions containing a large number of events, where a small systematic uncertainty can have an effect on the total yields as big as the expected signal contribution.

[2] Alessandro Manzoni, "I promessi sposi", Chapter 1 (1840).

The top quark mass is fixed to the value of $m_t = 172.5$ GeV and the EvtGen program is used to model the properties of bottom and charm hadron decays for all the non-SHERPA samples.

5.2.1 Signal samples

The $t\bar{t}H$ signal process is described via simulated samples produced with MADGRAPH5_aMC@NLO, for the generation of the hard-scatter event at NLO accuracy in QCD, interfaced with PYTHIA8 for the simulation of the parton shower (PS) and hadronization processes.

The decay of the top quarks is done with the MADSPIN [165] software in such a way that the spin correlations are preserved and the Higgs boson mass in the simulation is set to be $m_H = 125$ GeV, with all of the decay modes considered.

The signal cross section, $\sigma_{t\bar{t}H}$, is equal to 507^{+35}_{-50} fb, taken from calculations up to NLO in QCD and including NLO electroweak corrections [166–171].

5.2.2 $t\bar{t}$ + jets background modelling

The nominal $t\bar{t}$ + jets process is generated inclusively in all its subcomponents, namely $t\bar{t} + \geq 1b$, $t\bar{t} + \geq 1c$ and $t\bar{t} + light$, using POWHEG interfaced with PYTHIA8. Given that these subcomponents populate different regions of the selected phase space, the inclusive $t\bar{t}$ + jets sample is subdivided into these three main components, which are treated as separate samples.

The categorization is done according to the flavour of additional particle jets not originating from the $t\bar{t}$ system. Particle jets are reconstructed with the same anti-k_t algorithm ($R = 0.4$) as calorimeter jets, but by clustering stable truth particles[3], with the exclusion of muons and neutrinos. The kinematic selection for particle jets requires them to have a p_{T} greater than 15 GeV and $|\eta|$ less than 2.5.

If a jet is matched to exactly one b-hadron with $p_{\mathrm{T}} > 5$ GeV, the jet is labelled as single b-jet, whereas jets matched with more than one b-hadron are labelled B-jets, such as jets originating from gluon splitting

[3] Particles with a mean lifetime $\tau > 3 \cdot 10^{-11}$ s are considered as stable particles, as they are able to travel through the detector before decaying.

into a $b\bar{b}$ pair with small angular separation. Single c- and C-jets are defined in a similar way, but considering jets that are not already matched to one or more b-hadrons.

Events that contain at least one single b- or B-jet, excluding the ones coming from the top or W boson decays, are labelled as $t\bar{t} + \geq 1b$; while events with no extra b-jets but with at least one additional c-jet are labelled $t\bar{t} + \geq 1c$. Events labelled as either $t\bar{t} + \geq 1b$ or $t\bar{t} + \geq 1c$ are generically referred to as $t\bar{t}$+HF (HF stands for "heavy flavour"), whereas events without heavy flavour jets are labelled as $t\bar{t} + light$.

A finer classification is also provided: if an event has exactly two single b-jets, the event is labelled $t\bar{t} + b\bar{b}$, those with one single b-jet are called $t\bar{t} + b$ and those with exactly one B-jet are labelled $t\bar{t} + B$ events. Finally, remaining events enter in the $t\bar{t} + \geq 3b$ events. Events with additional b-jets coming from multi-parton interactions (MPI) or final-state radiation (FSR), i.e. originated from gluon radiation from the top quark decay products are considered in a separate category.

A second $t\bar{t}$ sample, which has a great impact on the flow of the analysis, is represented by the SHERPA + OPENLOOPS NLO $t\bar{t} + b\bar{b}$ sample [79, 147], referred in the following as SHERPA4F, as only the lightest four flavour quarks are considered massless. It represents the state-of-the-art of the theoretical knowledge of the $t\bar{t} + b\bar{b}$ process and is expected to provide the most accurate estimate of this process. This sample is used to reweight the fractions of the various subcategories of the $t\bar{t} + \geq 1b$ background as predicted by POWHEG+PYTHIA8, in order to improve its already good description of observed data.

This is possible because the description of the kinematics of the two additional b-jets is done at the NLO precision in QCD, taking the b-quark mass into account. As a matter of fact, considering massive b-quarks and massless light-quarks and gluons affects the balance between the various $t\bar{t}$ + jets production modes: the gluon splitting $g \to b\bar{b}$ dominates over the production of two b-quarks in the initial state.

In Figure 5.3 is shown a comparison of the predicted fractions of the various sub-categories of the $t\bar{t} + \geq 1b$ for the POWHEG+PYTHIA8 and SHERPA4F samples. The $t\bar{t} + b$ MPI/FSR sub-category, accounting for 10% of the events in POWHEG+PYTHIA8 $t\bar{t} + \geq 1b$ sample, is not present in the $t\bar{t} + b\bar{b}$ SHERPA4F sample and therefore is not scaled. The

uncertainties on the SHERPA4F prediction is derived by varying tune parameters, renormalization and factorization scales, as well as PDF sets for the SHERPA4F sample. A detailed source of the uncertainties entering the the band is given in Section 7 of Ref. [172].

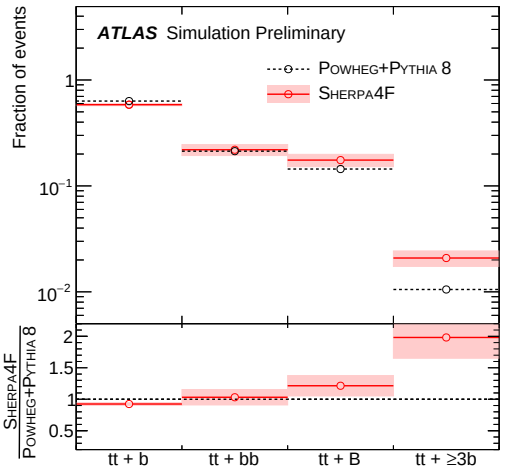


Figure 5.3: Comparison of the relative predicted fractions of the $t\bar{t}+b$, $t\bar{t}+b\bar{b}$, $t\bar{t}+B$ and $t\bar{t}+\geq 3b$ sub-categories, before any event selection, for the POWHEG+PYTHIA8 sample and the SHERPA4F. The fractions are normalized to the sum of the four contributions present in both generators, i.e. without considering the $t\bar{t}+b$ (MPI/FSR) sub-category as part of the total. The uncertainty band is derived with the procedure described in Section 7 of Ref. [172].

Four additional $t\bar{t}$ samples are generated to assess the modelling of the $t\bar{t}$ system. Three of them, namely the POWHEG+Herwig7 and the two POWHEG+PYTHIA8 with increased and reduced radiation in the final state, have already been described in Section 4.2.1 and they will not be described again. A sample generated with SHERPA and interfaced with OPENLOOPS, which considers the bottom-quarks massless, referred to as SHERPA5F in the following, is used to assess the matrix element generator. It should not be confused with the SHERPA4F sample used to reweight the fraction of $t\bar{t}+\geq 1b$.

5.2.3 Other backgrounds

Other processes rather than $t\bar{t}$ + jets can enter in the analysis regions with different yields depending on the jet and b-jet multiplicity of the region. Such background processes are taken from simulation with corrections applied to them, with the exception of the fakes and non-prompt lepton contribution, which in the SL channel is estimated using the data-driven Matrix Method described in Section 4.2.1. For the paper analysis, in the three most sensitive SR in the SL channel, the expected fake lepton background represents a minor contribution, compatible with zero and hence neglected. In the DIL channel this background is estimated from simulation, but normalized to data in a dedicated same-sign lepton region.

The single top, W/Z+jets and the diboson backgrounds are estimated using the same MC samples described in Section 4.2.1. For Z+jets events, an additional correction is applied to the normalization of the heavy-flavour component by scaling up the contribution by a factor 1.3, extracted from dedicated control regions with a definition close to the signal regions, but requiring the two leptons with opposite charge and same flavour to have an invariant mass close to the Z mass.

Samples of $t\bar{t}W$ and $t\bar{t}Z$, referred collectively as $t\bar{t}V$, are generated using MADGRAPH5_aMC@NLO interfaced with PYTHIA8.

The production of four top quarks in the final state, $t\bar{t}t\bar{t}$, and the production of a $t\bar{t}$ pair in association with a W boson pair, $t\bar{t}WW$, was generated with MADGRAPH5_aMC@NLO with LO accuracy and interfaced with PYTHIA8. On the other hand, tZ events were still produced with MADGRAPH5_aMC@NLO with LO accuracy, but interfaced with the PYTHIA6. Finally, MADGRAPH5_aMC@NLO samples interfaced with PYTHIA8 are used to describe the tZW process at NLO accuracy.

The associated production of the Higgs boson with a single top quark is very small in the SM, but is nevertheless included in the analysis and treated as background. The tWH production is modelled via samples generated with MADGRAPH5_aMC@NLO interfaced with Herwig++, whereas samples describing the $tHqb$ production mode were produced with MADGRAPH5_aMC@NLO interfaced with PYTHIA8 at LO accuracy. Other Higgs boson production modes are negligible and hence not considered in the analysis.

5.3 Analysis strategy for ICHEP

A first version of the analysis using only 13.2 fb^{-1} collected during 2015 and 2016 was presented at the ICHEP conference in Chicago, later referred as the "ICHEP analysis". Most of the MC samples and strategies described in the previous sections were used also in this version of the analysis, with the most notable difference that the nominal $t\bar{t}$ sample was generated with POWHEG interfaced with PYTHIA6. Nevertheless, the discussion and the conclusions presented in this section are not affected by these considerations.

Given that the techniques developed and described in the following apply to the single-lepton channel only, the dilepton channel will not be described.

5.3.1 Event selection and categorization

The selection requires exactly one isolated lepton. Events must pass the same single-lepton triggers used for the calibration analysis of the Jet Vertex Charge.

Events are required to have at least four jets with $p_{\mathrm{T}} > 25$ GeV and $|\eta| < 2.5$ and at least two of them have to be b-tagged using the MV2c10 algorithm. The chosen working point corresponds to an efficiency of 70% for the selection of b-jets in $t\bar{t}$ simulated events.

Events are later classified into exclusive regions based on the number of jets and b-jets. A region with m jets and n b-jets is referenced as (mj, nb). Regions are defined by having exactly four, five or at least six jets and exactly two, three and at least four b-jets. Figure 5.4a shows the background composition in the various regions of the analysis. The composition is dominated by $t\bar{t}$ events in all regions, with $t\bar{t}$+HF events becoming more important in regions with a higher jet and b-jet multiplicity. In Figure 5.4b are shown the $S/\sqrt{B}$ and S/B ratios in each of analysis region. The red (5j, $\geq$ 4b), ($\geq$ 6j, 3b) and ($\geq$ 6j, $\geq$ 4b) regions are treated as signal regions, whilst the remaining ones are considered as control regions.

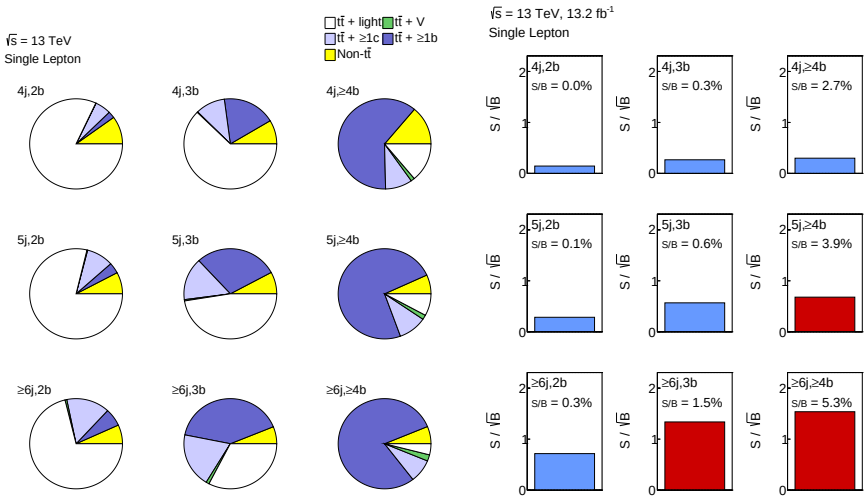


Figure 5.4: Pie charts with the fractional predicted contributions of the various backgrounds to the total background prediction (left) and the $S/\sqrt{B}$ and S/B ratios (right) for each of the analysis regions for the single-lepton channel. Each row corresponds to a different jet multiplicity and each column corresponds to a different b-jet multiplicity. Signal regions are coloured in red while control regions in blue.

5.3.2 Reconstruction BDT

In this version of the analysis, the reconstruction of the final state in the SRs was performed only by means of BDTs, known as "reconstruction BDT" (recoBDT), trained to match the jets to the corresponding partons in a $t\bar{t}H(b\bar{b})$ process.

For training purposes, only the $t\bar{t}H$ simulated sample was used. The training signal is defined as the correct permutation of the objects for which the assignment matches the truth record, whereas all the other permutations are considered as background.

Both topological and kinematic variables are used as input for the recoBDT, such as the mass of the leptonic and hadronic top, as well as

for the W system or ΔR distance between pairs of objects; the full list of variables can be found in Appendix A. For each permutation of the final state objects, these variables are computed and the corresponding BDT output is evaluated: the permutation with the highest BDT output discriminant is chosen to be the permutation representative of the parton-to-jet assignment. In order to reduce the number of possible permutations, as well as the computing time, the b-tagged jets can only be in the position where a b-jet is expected, i.e. b-jets from the top quark decays or from the $H \to b\bar{b}$ decay.

Two versions of the recoBDT were built and both of them were employed in the final analysis: the version with and without variables sensitive to the Higgs boson system. This was done so that the presence of variables related to the Higgs system does not bias the output towards permutations that have an expected Higgs mass peak around 125 GeV, in particular when applied to the $t\bar{t}+b\bar{b}$ background. In this way, it is possible to construct variables related to the Higgs system with the best permutation given by the recoBDT and use them as input for the second stage of the event classification.

The W boson decaying into a pair of quarks, W_{had}, is reconstructed by pairing two non b-tagged jets in the event. If there are fewer than two light jets in the event, one b-jet is used for its reconstruction, except in the $(5\text{j}, \geq 4\text{b})$ region, where the W_{had} is not reconstructed, as the missing jet is usually the sub-leading one from the W decay.

The missing energy is used as a proxy for the reconstruction of the neutrino four-momentum. The component of the momentum along the z direction is obtained by imposing that the invariant mass of the lepton and the neutrino equals the W boson mass. Both real solutions of the resulting quadratic equation are considered; in case there is no real solution, the discriminant is set to zero.

The two top quarks are reconstructed by matching a W with one b-jet in the event. Finally the Higgs boson candidate is built with the remaining b-jets in the event. In the $(\geq 6\text{j}, 3\text{b})$ region, one b-jet can be used to either build the Higgs candidate or the top candidates, whereas in the region with five selected jets, the hadronic top is reconstructed by using a b- and a light-jet.

In the $(\geq 6\text{j}, \geq 4\text{b})$ signal region the maximum efficiency achieved

in matching correctly all the objects is 13.7% when using the recoBDT with Higgs related variables as input and 10.4% without those variables. These values have to be compared to a maximum theoretical efficiency of 38%, which represent the fraction of events entering in this region for which all of the $t\bar{t}H(b\bar{b})$ decay objects pass the event selection. The matching efficiencies for all the objects and their combinations are shown in Figure 5.7.

The big difference between the observed efficiency of 13.7% and the potential 38% has various reasons. Among them, there is definitely the complexity of the final state that has to be reconstructed. In particular, the presence of jets from additional parton emission in the final state increases the number of possible permutations, so that it's not uncommon to find the correct permutation as the one with the second or third highest BDT output.

The reconstructed Higgs boson invariant mass in the most sensitive signal region can be seen in Figure 5.5. The distribution for the $t\bar{t}H(b\bar{b})$ signal events is shown by the blue histogram, while the distribution with the correct jets assignment to the Higgs boson is shown by the filled histogram.

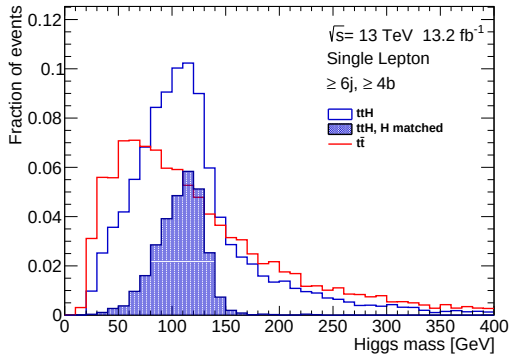


Figure 5.5: Higgs boson invariant mass from the reconstruction BDT trained without using Higgs-related variables in the ($\geq$ 6j, $\geq$ 4b) region. All signal events are shown in blue, while those with the correct jets assigned to the Higgs boson are shown in the solid blue histogram. The $t\bar{t}$ background, consisting primarily of $t\bar{t} + \geq 1b$, is shown in red.

5.3.3 Use of the Jet Vertex Charge in the reconstruction

The Jet Vertex Charge tagger was developed with the goal of improving the reconstruction of the $t\bar{t}H(b\bar{b})$ final state. The main concept is the possibility to use the parton electric charge to resolve ambiguities in the b-jet assignment.

In fact, by drawing the corresponding Feynman diagram it is easy to see that the b-quark coming from the leptonic top has an opposite charge with respect to the charge of the lepton, whereas the opposite happens for the b-quark in the hadronic top hemisphere. Additionally, the b-quarks from the Higgs decay must have opposite charge with respect to each other.

Therefore, probabilities can be assigned to the b_{lep} and b_{had} jets of being of opposite or same sign with respect to the charge of the lepton by using the correlation between the λ_{JVC} discriminant of the jet and the charge of the lepton as:

$$\begin{aligned} P(b_{\text{lep}}) &= \begin{cases} P(b^-) & \text{if } Q_{\text{lep}} > 0 \\ P(b^+) & \text{if } Q_{\text{lep}} < 0 \end{cases} \\ P(b_{\text{had}}) &= \begin{cases} P(b^-) & \text{if } Q_{\text{lep}} < 0 \\ P(b^+) & \text{if } Q_{\text{lep}} > 0 \end{cases} \end{aligned} \tag{5.1}$$

and with the same logic the probability that the two b-jets associated with the Higgs decay have opposite charge can be constructed as:

$$P(\text{Higgs}) = P(b_1^+)P(b_2^-) + P(b_1^-)P(b_2^+) \tag{5.2}$$

where the probability of being a positive or negative b-jet, given its λ_{JVC} discriminant, is defined as:

$$\begin{aligned} P(b^+|\lambda_{\text{JVC}}) &= \frac{e^{\lambda_{\text{JVC}}}}{1+e^{\lambda_{\text{JVC}}}} \\ P(b^-|\lambda_{\text{JVC}}) &= \frac{1}{1+e^{\lambda_{\text{JVC}}}} \end{aligned} \tag{5.3}$$

These variables are built on top of each permutation defined for the default recoBDT; the new reconstruction BDT trained with their inclu-

sion will be referred to as recoJVC. Their distributions are shown in Figure 5.6 for the region ($\geq$ 6j, $\geq$ 4b), whereas the same variables for all the SR are presented in Appendix B.1. For the recoJVC, the signal permutation (blue histogram) and the background permutations (red histogram) are defined in the same manner as for the default recoBDT and the permutation with the highest BDT score is chosen as the permutation that matches the objects to the truth counterparts.

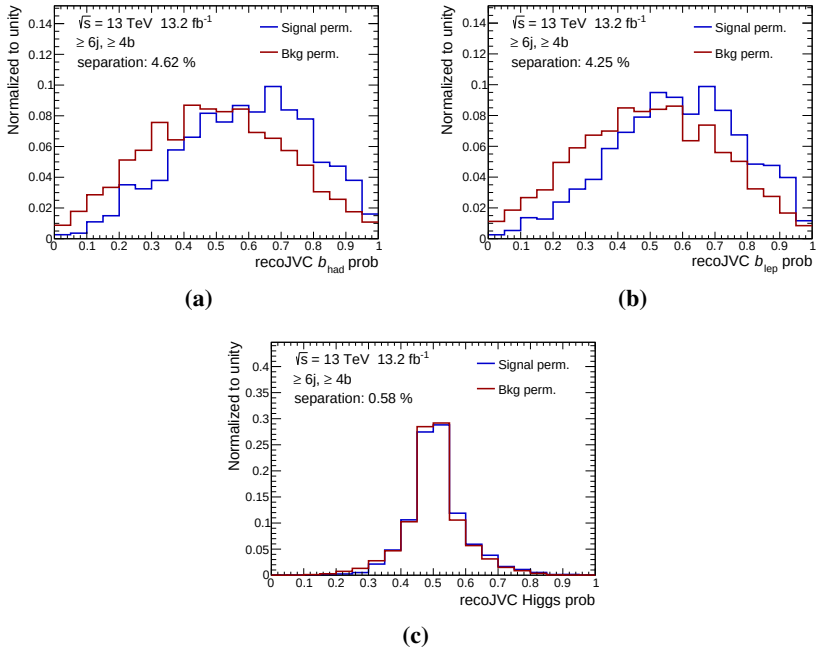


Figure 5.6: Distribution for the b_{had} (top left), b_{lep} (top right) and Higgs (bottom) probability as defined by Eqs. (5.1) and (5.2) for the signal and background permutation in the ($\geq$ 6j, $\geq$ 4b) SR.

The discriminating power between signal and background for each input variable can be quantified by measuring the *separation*, that is the non-overlapping area of the two histograms, defined as:

$$S = \frac{1}{2} \sum_{\text{i} \in \text{ bins}} \frac{(N_i^{\text{sig}} - N_i^{\text{bkg}})^2}{N_i^{\text{sig}} + N_i^{\text{bkg}}} \tag{5.4}$$

By including these variables, the matching efficiencies of the reco-JVC improve, as shown in Figure 5.7. Three values are reported for each of the category: the efficiency for the default training (red), for the training with the inclusion of charge-variables (blue) and the trainings for which the truth charge information is used to replace the corresponding charge variables (green).

In the truth trainings, the charge-probabilities were replaced with the charge of the b-quark matched to the jet, hence those trainings are a way to assess the maximum achievable improvement that charge information can bring to recoJVC: they serve as a reference for the maximum possible achievable improvement for the ideal case of a Jet Vertex Charge algorithm with optimal separation.

The efficiency for matching all the objects in the $(\geq 6\text{j}, \geq 4\text{b})$ SR equals 13.7% when using Higgs related variables, increases to 14.8% when using charge-probabilities and is 18.2% when training with truth charge information. Overall, a similar increase in the matching efficiency is seen for the various objects and their combination. A similar situation is observed for the trainings without the Higgs variables.

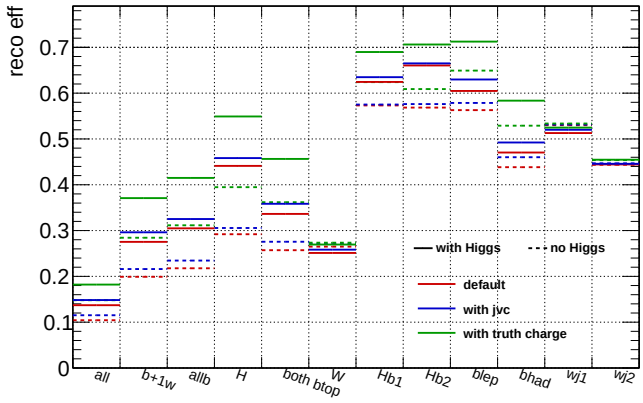


Figure 5.7: Matching efficiency of various objects and their combination in the $(\geq 6\text{j}, \geq 4\text{b})$ signal region for the recoJVC in the version with (solid line) and without Higgs-related variables (dashed line). The default training (red), with charge-variables (blue) and with truth charge (green) are overlaid.

5.3.4 Classification BDT

In order to classify events as more signal- or background-like, a second layer of multivariate algorithms is trained in each of the signal regions. This second layer is called "classification BDT" (classBDT) and combines information from the output of recoBDT with kinematic variables that describe the event topology.

Simulated $t\bar{t}H(b\bar{b})$ events are used as signal and $t\bar{t}$ as background. The choice of the input variables is made independently in each of the three signal regions, given the important differences in jet and b-jet multiplicities. The variables used in these BDTs are listed in Appendix C.

In order to test the effect of the inclusion of Jet Vertex Charge inputs in the final state reconstruction, alternative classBDTs are trained, for which the input variables coming from the recoBDT are replaced by the corresponding ones obtained from the (truth) recoJVC. Furthermore, the hyperparameters of the various classBDTs were reoptimized to profit from the additional input variables.

The output discriminants of the three trainings for the three SRs of the analysis are shown in Figure 5.8.

In the most powerful SR the improvement in separation due to using recoJVC inputs is marginal compared to the default classBDT, in spite of a possible few percent gain observed with the truth trainings.

In the $(\geq 6\text{j}, 3\text{b})$ region, both the improvements given by recoJVC and the truth charge are marginal. This is due to the fact that one b-jet is missing in the final state, therefore a full reconstruction of the final state is impossible.

On the other hand, in the $(5\text{j}, \geq 4\text{b})$ SR, there is a worsening of the separation in the final classification BDT output between $t\bar{t}H(b\bar{b})$ and $t\bar{t}$. The impossibility to have a full reconstruction of the final state is playing an important role in this case as well. In particular, the performance degrades with respect to the default classBDT.

A metric more robust against the choice of the binning and statistical fluctuations is shown on the plots as well: the area under the ROC curve (AUC). The performance of a generic classifier can be evaluated by looking at the background rejection rate as a function of the signal efficiencies, known as Receiver Operating Characteristic (ROC) curve, and the area under this curve is a common measure of the performance

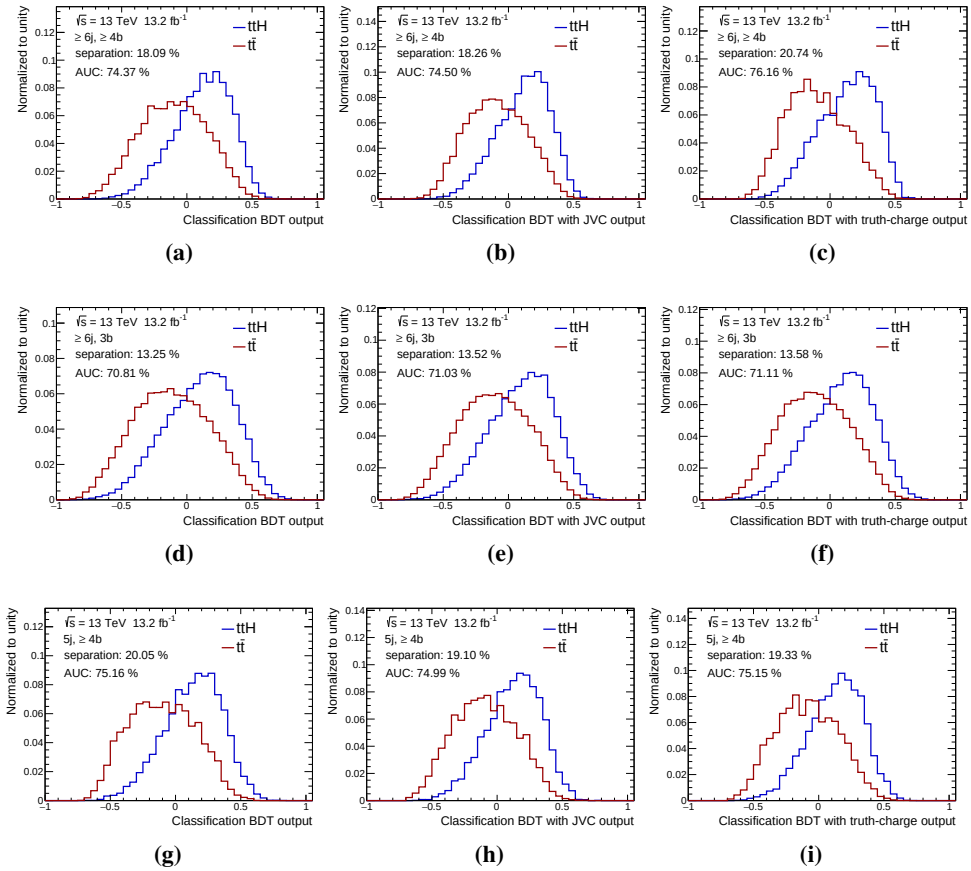


Figure 5.8: Output discriminant of the classification BDT in ($\geq$ 6j, $\geq$ 4b) region (top), ($\geq$ 6j, 3b) (centre) and (5j, $\geq$ 4b) (bottom) in the SL channel trained using the default setup (left), including JVC (centre) and with the truth-charge (right).

of a classifier. Nevertheless, the conclusions remain identical.

The lack of improvement, in spite of a general increase in the matching efficiency of recoJVC, can be traced back to a small increase in the separation between signal $t\bar{t}H(b\bar{b})$ and the $t\bar{t}$ + jets background for the variables built from the output of the reconstruction BDT. Two repre-

sentative variables are shown in Figure 5.9, namely the distribution of the BDT output of the best permutation and the invariant mass of the Higgs candidate, while the full list of output variables of recoJVC is reported in Appendix B.2 for the three SRs.

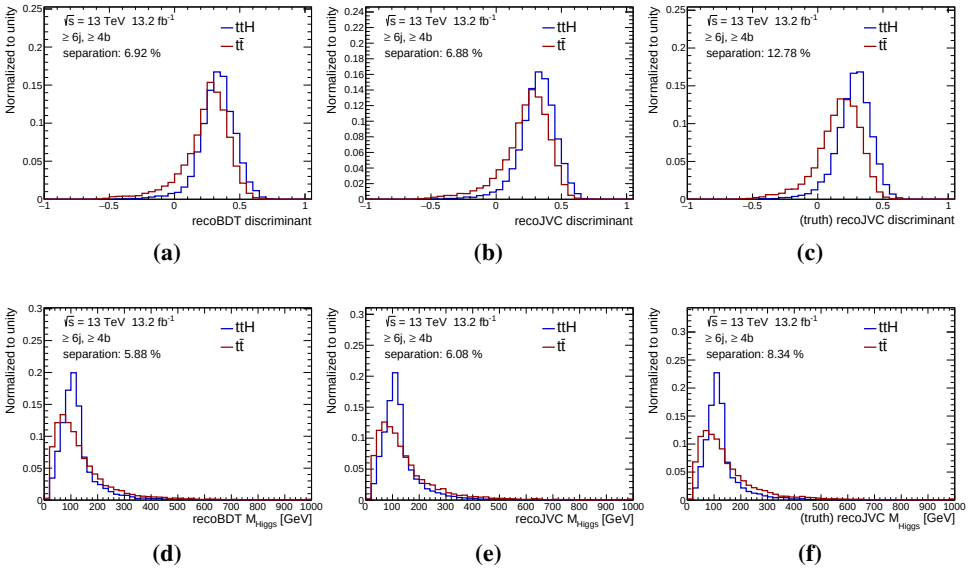


Figure 5.9: Output discriminant of the recoBDT (top) and reconstructed Higgs mass (bottom) coming from the recoBDT for the "nominal" training (left), training with JVC (centre) and truth charge (right) for ($\geq$ 6j, $\geq$ 4b) region in the SL channel.

The reason for a similar separation between the three trainings resides in the fact that most important background, $t\bar{t} + \geq 1b$, has the same charge signature as the signal $t\bar{t}H(b\bar{b})$. In Table 5.1 the separation between the $t\bar{t}H(b\bar{b})$ signal and the three $t\bar{t}$ subcomponents is presented for output BDT and the three trainings. It is clear that the separation between the default training and the JVC ones comes mostly from the $t\bar{t} + light$ component, which is of minor importance in the SR.

For these reasons, it was decided not to use it in the final analysis.

Table 5.1: Separation for the three flavours of $t\bar{t}$ + jets background for the three recoBDT in the ($\geq$ 6j, $\geq$ 4b) region for the output BDT variable.

	default	JVC	truth
$t\bar{t} + \geq 1b$	7.25 %	7.14 %	12.96 %
$t\bar{t} + \geq 1c$	6.22 %	6.23 %	13.90 %
$t\bar{t} + light$	8.68 %	9.23 %	12.64 %

5.3.5 Disentangling $t\bar{t}$+HF jets from $t\bar{t} + light$ jets

A limiting factor in the final signal extraction comes from the large uncertainties associated with the $t\bar{t}$+HF background and by the impossibility to disentangle its contribution from the $t\bar{t}H(b\bar{b})$ signal.

Having a region pure in $t\bar{t}$+HF that enters the likelihood fit can be beneficial for the analysis; in fact, this region can be used by the fit to improve the knowledge of the backgrounds and, as a consequence, reduce the associated modelling uncertainties. The improvement is possible by exploiting the shape differences among the backgrounds in order to disentangle degenerate systematics, i.e. uncertainties that have a similar if not identical effect in shape and acceptance variations in the total background prediction.

It can be understood by thinking that, once one of the background components is known, it will be extrapolated and fixed in all the other regions of the analysis, effectively removing some degrees of freedom so that the remaining background uncertainties can be determined more precisely. It is, in all respects, analogous to making a measurement *in-situ*.

For this reason an additional BDT, named HFBDT, is trained in the (5j, 3b) CR to discriminate $t\bar{t}$+HF from $t\bar{t} + light$, with the aim of defining a region pure in $t\bar{t}$+HF. This region is chosen as it offers low signal contamination combined with a good fraction of $t\bar{t}$+HF events (about 44% of the total expected background) and, equally important, with reasonable number of events to train the BDT.

The difference between $t\bar{t} + light$ and $t\bar{t}$+HF events in this region lies in the origin of the third b-tagged jet: in a real $t\bar{t} + light$ event, the third

b-tagged jet is the result of a mis-tag of a light or charm-jet. On the contrary, additional HF partons in a $t\bar{t}$+HF event coming from ISR or produced in the hard scatter are likely to be *b*-tagged.

For this reason, the third *b*-tagged jet in a $t\bar{t}+light$ event is likely to come from a mis-tag of the *c*-quark from the *W* boson decay, given that the *W* boson decay products contain a *c* quark in almost 30% of the cases and the mis-tag rate is higher compared to the one of a light-quark. Hence, differences between the $t\bar{t}+light$ and $t\bar{t}$+HF come from the kinematics of the *b*-tagged jet pair, as well as in the possibility to reconstruct the W_{had} mass from the two non *b*-tagged jets. These differences are used to build the input variables for the HFBDT.

Choice of the input variables for HFBDT

The final set of input variables is determined by first defining a pool of candidate variables and second by selecting the most powerful ones, as it is generally preferable to have a limited number of inputs. The ranking of the input variables is often used as guidance; variables were ranked based on how often one variable is used in a node of the BDT and the gain in separation obtained after the node-splitting. Furthermore, variables with high correlations and comparable performance can be either substituted by a combination of the two or simply replaced by just one of the two.

An iterative procedure is used: the least ranked variable is removed, a new training is performed and its performance evaluated. The process is interrupted when a balance between performance of the BDT and its complexity is reached, with the complexity being related both to the number of input variables to validate and their correlations.

Two figures of merit were used in order to quantify the discrimination power of the BDT output: the separation, as defined by Eq. (5.4), and the area under the ROC curve. Both metrics are plotted as a function of the number of input variables in Figure 5.10. A plateau is reached with the set of eleven variables. The full list of input variables is presented in Table 5.2, together with a brief explanation of their meaning. The separation between $t\bar{t}$+HF and $t\bar{t}+light$ flavour for all the input variables is shown in Figure D.1 of Appendix D.

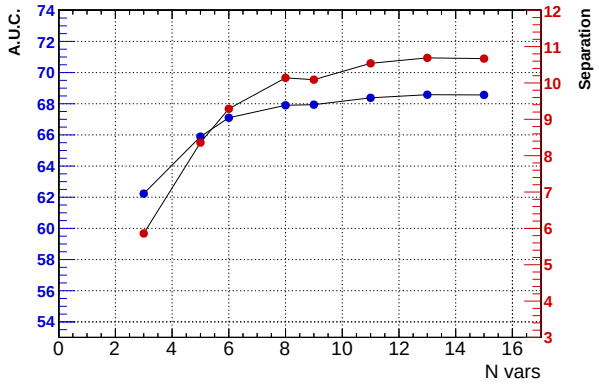


Figure 5.10: Separation and area under the ROC curve (AUC) as a function of the number of input variables used in the HFBDT.

Table 5.2: List of the input variables employed by the HFBDT.

Variable	Definition
$m_{bb}^{\text{min }\Delta R}$	Invariant mass of the two b-tagged jets with the smallest ΔR
$m_{bb}^{\text{max }p_T}$	Invariant mass of the two b-tagged jets with the largest vector sum p_T
$m_{uu}^{\text{min }\Delta R}$	Invariant mass of the two non b-tagged jets with the smallest ΔR
$m_{jj}^{\text{min }\Delta R}$	Invariant mass of any two jets with the smallest ΔR
$\Delta R_{uu}^{\text{min }\Delta R}$	ΔR of the combination of the two non b-tagged jets with the smallest ΔR
$\Delta R_{\text{lep}-bb}^{\text{min }\Delta R}$	ΔR between the lepton and the combination of two b-tagged jets with the smallest ΔR
$\Delta R_{bb}^{\text{avg}}$	Average ΔR for all b-tagged jet pairs
$m_{bb}^{\text{max M}}$	Invariant mass of the two b-tagged jets with the largest invariant mass
$m_{jjj}^{\text{max }p_T}$	Invariant mass of any three jets with the largest vector sum p_T
$m_{bj}^{\text{mass }W}$	Invariant mass of a b-tagged jet and any jet with the invariant mass closest to the W mass
$\Delta R_{bj}^{\text{mass }W}$	ΔR between the combination of a b-tagged jet and any jet with the invariant mass closest to the W mass

HFBDT validation

Since BDTs are prone to suffer from *overtraining*, sanity checks need to be performed in order to validate the training procedure and gain confidence in its robustness.

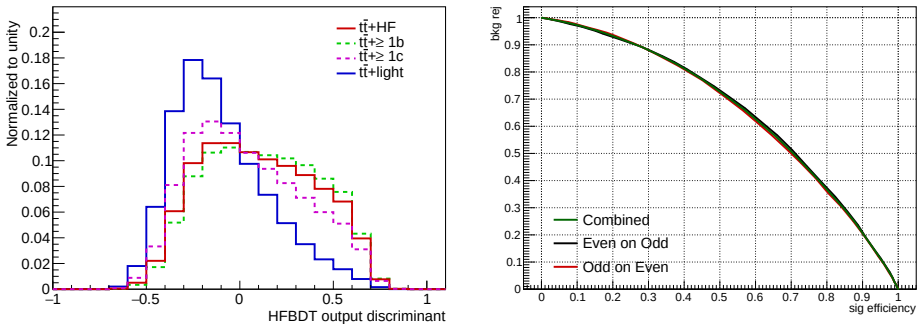
If the size of the training sample is not sufficient, the MVA algorithm may have specialized excessively and picked up specific features due to statistical fluctuations, rather than physical differences in the input variables. The performance obtained on the training sample has to be reproducible in an independent, unseen new sample, called test sample; therefore overtraining is visible when the performance measured on an independent test sample is not equal to the one expected from the training sample.

Furthermore, the available MC statistics can often represent a problem and a limiting factor when using MVA techniques, as the initial sample has to be divided into a training and a testing sample, effectively reducing the statistics available.

The k-fold cross validation is a way to both check for overtraining and recover the lost statistics. The initial sample is divided into k subsamples (folds) in a random way, one of which is used for the testing of the algorithm and the other $k-1$ subsamples are used for the training itself. This process is then repeated k times and the k test results can be averaged to get a better estimator of the performance of the trained model. The advantage of this method is that the whole dataset is used for both the training and the validation of the MVA.

A 2-fold validation is used as a cross-check. In Figure 5.11a the BDT discriminant distribution is shown, while the ROC curves for the two samples (called Even and Odd) are shown in Figure 5.11b as well as the combined ROC curve for the combined sample.

The validation of the input variables of an MVA is a step of particular importance; not only the MC has to provide a good description of the variables, but also their correlation has to be described accurately, in order to have a reliable output discriminant. In fact, even if the BDT approach, as opposed to a traditional cut-based analysis, allows to exploit further the correlations among the input variables, it is good practice to replace a pair of highly correlated variables that have similar separation with a single variable.



(a) **(b)**

Figure 5.11: The distribution of the final discriminant for signal and background for the HFBDT (left) and the ROC curves for the 2-fold validation (Even on Odd and Odd on Even) as well as the combined ROC curve for the combined sample for the HFBDT (right).

The correlation can be measured, at first order, with the *linear correlation coefficient*, ρ, defined as:

$$\rho(x,y) = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{\sum_i (x_i - \mu_x)\cdot(y_i - \mu_y)}{\sqrt{\sum_i (x_i - \mu_x)^2}\sqrt{\sum_i (y_i - \mu_y)^2}} \tag{5.5}$$

where x and y are the two variables under investigation, cov represents the covariance matrix, $\mu_i(\sigma_i)$ is the mean (RMS) of variable i and the sum is done over all the events passing the selection.

It should be noted that independent variables will have ρ=0, but the inverse does not necessarily hold.

The following steps have been used to validate the input variables:

1. ensure that all the variables are correctly modelled by the MC simulation and eventually remove variables poorly modelled;

2. plot correlation matrices for MC and data, as well as the linear correlation coefficient for every pair of variables whose correlation in MC and data differ by more the $|\Delta\rho| > 10\%$;

3. plot the distribution of ρ for every pair of variables with significant correlation, $|\rho| > 20\%$.

Data/MC comparisons of all the input variables are shown in Figure 5.12. A good description of data is found in the Monte Carlo simulated samples, as the ratio of the two at the bottom of the plot is well within the uncertainty band and does not show any trend.

The linear correlation matrices for MC and data are displayed in Figure D.2. No difference larger than 10% is observed, indicating that the MC is capable of describing data with good agreement. Furthermore, a good data/MC agreement is found in all linear correlation variables that are larger than 20% in MC, as can be seen in Figure D.3 and D.4.

A discrepancy in the observed data and predicted MC yields is observed in all the regions rich in $t\bar{t}$+HF, including the (5j, 3b) region. This excess is compatible with the prediction given the large uncertainties associated with the $t\bar{t}$+HF production [162, 163]. For this reason, normalization factors are used to scale the $t\bar{t} + \geq 1b$ and $t\bar{t} + \geq 1c$, as detailed in the next section, and the contributions from the $t\bar{t}$ + jets subcomponents have been scaled up in all the plots relative to the HFBDT to reflect the yields post-fit, in order to better compare the shapes.

Impact on the analysis

In order to estimate the impact of the HFBDT, comparisons of the final analysis results have been carried out between the nominal setup of the analysis and a modified one in which the discriminating variable in the (5j, 3b) control region is the HFBDT output. Such comparisons have been done on both an Asimov and the real, blinded dataset, with a setup similar to the one used for the analysis described in Section 5.4. The main differences are represented by the definitions of the regions that enter in the likelihood fit and the details of the systematic model, which are irrelevant for the following discussion. The reader is referred to Section 5.5 for a detailed description of the Asimov dataset, as well as the likelihood method itself.

Since the energy required to produce the $t\bar{t}H$ signal is higher than the one needed to produce the $t\bar{t}$ background, signal events are expected to be on average more energetic and more central in the detector than

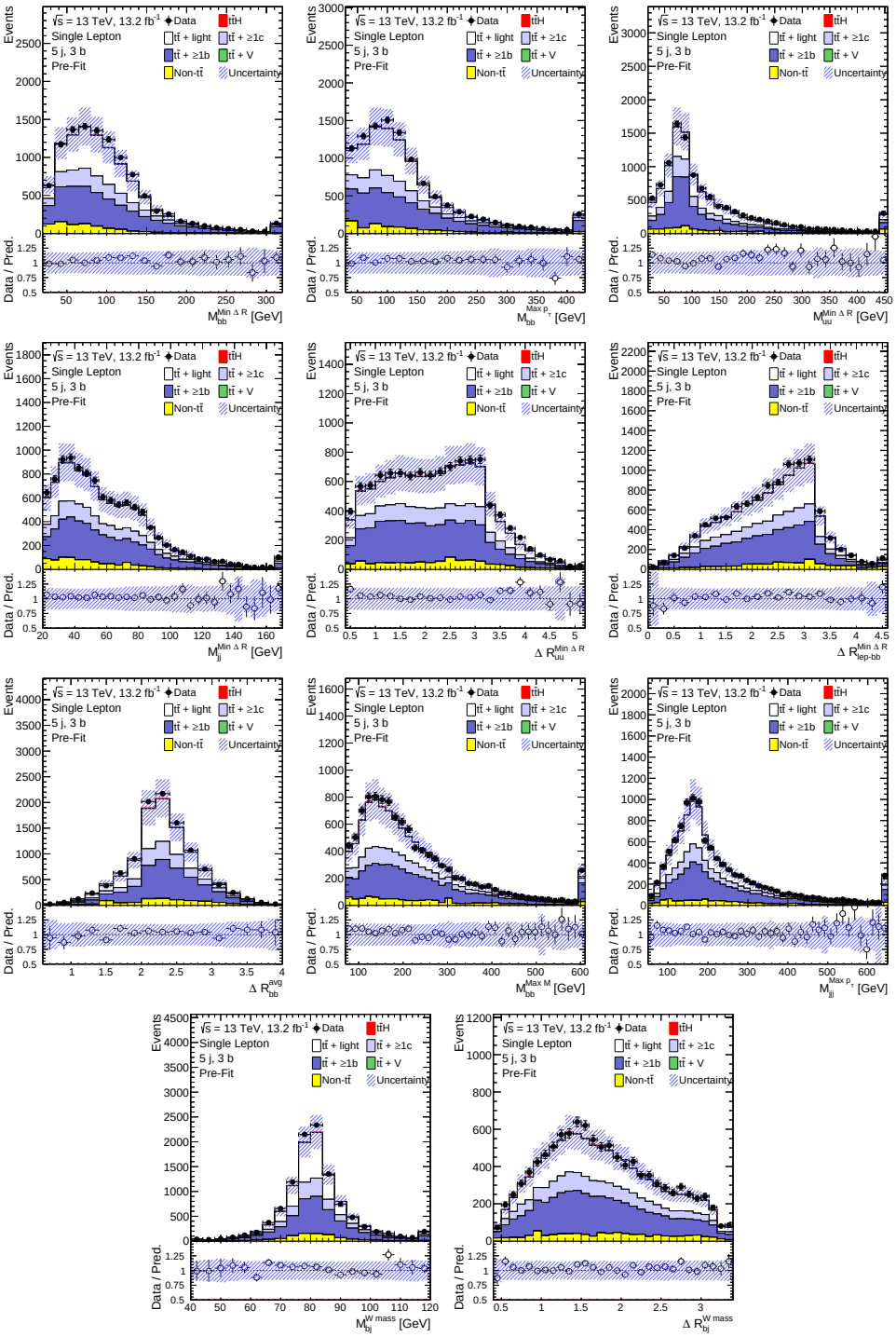
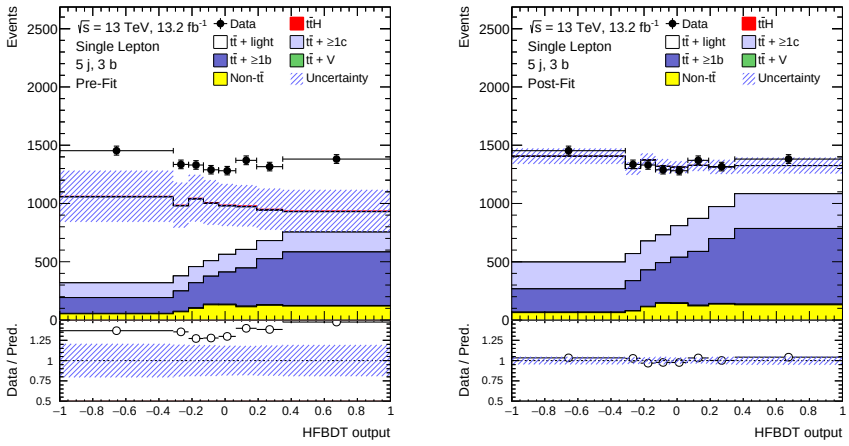


Figure 5.12: Data/MC plots of all the input variables to the HFBDT. The ratio panel at the bottom shows a good description of data by the simulation.

background ones. In the control regions the variable H_{T}, defined as the scalar sum of the p_{T} of all the jets in the event, is used as input in the likelihood fit, whereas the output of the classification BDT is used in the signal regions only in the Asimov fits. In these fits, the "Signal-plus-Background" hypothesis is used; whereas in fits on real blinded data the "Background-only" hypothesis is used, excluding bins for which $S/B > 5\%$. A floating factor μ, called signal strength, is used to scale the signal contribution and two normalization factors are used to scale the $t\bar{t} + \geq 1b$ and $t\bar{t} + \geq 1c$ contributions, called κ_{ttb} and κ_{ttc}. A fitted value of 1 indicates agreement with the SM or with the background prediction, respectively.

The fit is performed for the single-lepton channel only. Figure 5.13 shows the distribution of HFBDT discriminant pre- and post-fit under the background-only hypothesis. A good agreement is found post-fit, as the data/MC ratio lies within the uncertainty band, reduced considerably by the fitting procedure.



(a) **(b)**

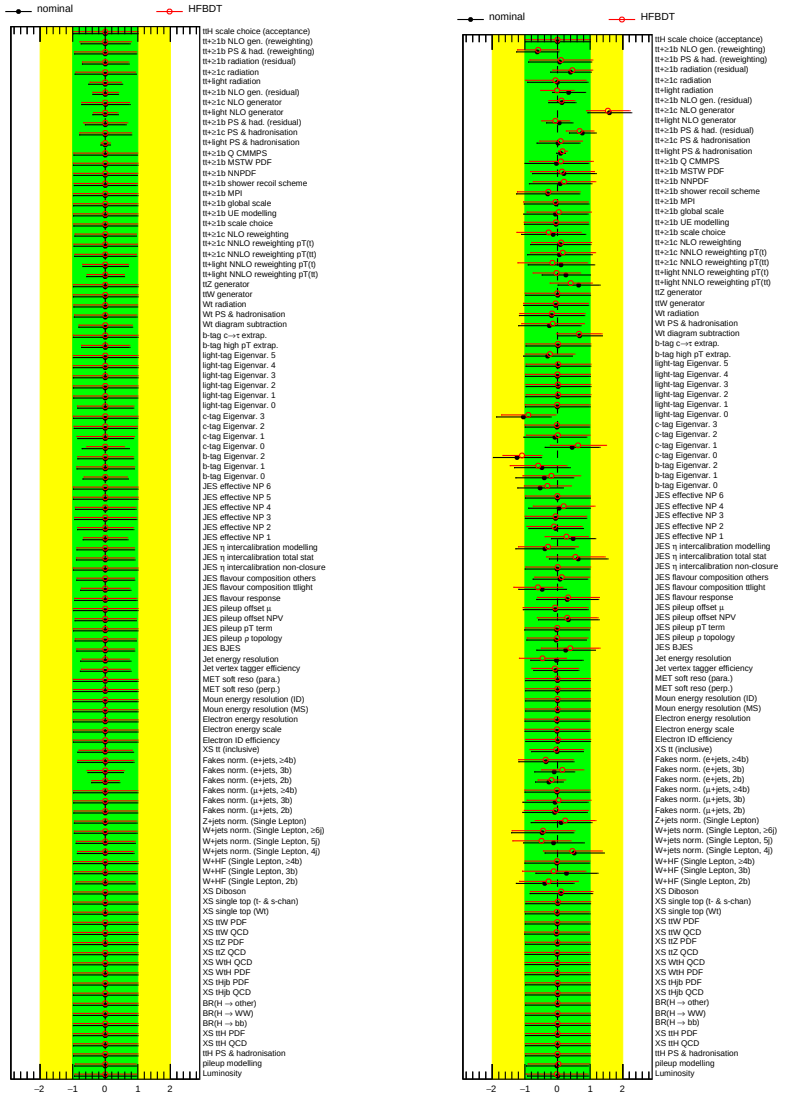
Figure 5.13: HFBDT output discriminant distribution pre-fit (left) and post-fit (right). The fit is performed under the background-only hypothesis.

The post-fit uncertainties on μ and the normalization factors are reported in Table 5.3 for both the Asimov fit and the blinded data fit.

Figure 5.14 shows the comparison of the nuisance parameters resulting from the both fit configurations. For the Asimov fit, no significant difference is visible; on the other hand, when fitting real data, some of the nuisance parametrization are pulled differently, but with a marginal impact. From the pull analysis, the fact that there is no reduction of the expected error on μ and the fact that κ_{ttb} and κ_{ttc} are compatible in both cases made it not necessary to employ an extra layer of complexity in the analysis, hence the HFBDT discrimination was not used.

Table 5.3: Comparison of the post-fit values for the signal strength μ and the normalization factors κ_{ttb} and κ_{ttc} are reported for the fit performed under the signal-plus-background hypothesis on the Asimov dataset and the background-only hypothesis on blinded data.

	Asimov		Data	
	nominal	HFBDT	nominal	HFBDT
μ	$1.00^{+0.95}_{-0.87}$	$1.00^{+0.95}_{-0.87}$	–	–
κ_{ttb}	$1.00^{+0.19}_{-0.17}$	$1.00^{+0.17}_{-0.16}$	$1.43^{+0.23}_{-0.20}$	$1.54^{+0.23}_{-0.21}$
κ_{ttc}	$1.00^{+0.60}_{-0.47}$	$1.00^{+0.60}_{-0.47}$	$1.28^{+0.75}_{-0.69}$	$1.40^{+0.77}_{-0.63}$



(a) **(b)**

Figure 5.14: Comparisons of the fitted nuisance parameters obtained with the nominal fit (black full circles) and a fit with the HFBDT in the (5j, 3b) CR (red open circles). The fit on the Asimov dataset under the signal-plus-background hypothesis is shown on the left and the fit on real data under the background-only hypothesis is on the right. In all the CR the H_{T} variable is used and the classBDT discriminants are employed in the SR only for the Asimov fit.

5.4 Strategy for paper analysis

This Section contains the final version of the analysis searching for the $t\bar{t}H(b\bar{b})$ process that uses the full 2015 and 2016 dataset, for a total of 36.1 fb^{-1} [172]. It underwent various changes with respect to the ICHEP analysis, with the most significant ones being the definition of a *boosted* signal region and the possibility to use the so-called *pseudo-continuous b-tagging*, which allowed the simultaneous use of the four calibrated working points (WP), 85%, 77%, 70% and 60%, that led to new definitions of signal and control regions.

The samples outlined from Sections 5.2.1 to 5.2.3 were used for this analysis.

5.4.1 Event selection and classification

As was the case for the ICHEP analysis, selected events are required to have exactly one (two) isolated leptons for the single-lepton (dilepton) channel. All events must pass single-lepton triggers, the same triggers used for the ICHEP and the Jet Vertex Charge calibration analysis.

Events in the single-lepton channel must contain exactly one lepton with $p_{\mathrm{T}} > 27$ GeV and no other leptons with $p_{\mathrm{T}} > 10$ GeV. In case an event contains two or more τ_{had} candidates[4], the event is removed, in order to avoid overlap of selected events with other $t\bar{t}H$ analyses containing τ_{had} in their final state.

Events enter the dilepton channel if they contain exactly two leptons with opposite electric charge. The leading lepton is required to have a $p_{\mathrm{T}} > 27$ GeV, while the subleading lepton p_{T} must be above 15 GeV in the ee channel or above 10 GeV in the $e\mu$ and $\mu\mu$ channels. In the same-flavour channels (ee and $\mu\mu$), the invariant mass of the dilepton system must be above 15 GeV and outside of the Z boson mass window 83 - 99 GeV. Dilepton events are vetoed if they contain one or more τ_{had} candidates, in order to be orthogonal with other $t\bar{t}H$ search channels.

If a top quark or the Higgs boson has a high transverse momentum,

[4] τ leptons decaying into hadrons (τ_{had}) are distinguished from jets using the track multiplicity and a multivariate discriminant based on the track collimation, jet substructure, and kinematic information [96].

it is said to be *boosted* and its decay products will be collimated. For the definition of the boosted region, small-R jets ($R = 0.4$) are reclustered [173] by being fed as input to the anti-k_t algorithm with a radius parameter of $R = 1.0$, resulting in a collection of large-R jets. Only large-R jets with an invariant mass greater than 50 GeV are considered for further selection criteria.

For the definition of the *boosted* signal region, boosted Higgs boson candidates are required to have $p_{\mathrm{T}} > 200$ GeV and contain at least two small-R jets, at least two of which have to be b-tagged with the 85% working point. If more than one boosted Higgs boson candidate is identified, the selected one corresponds to the candidate whose sum of the small-R jet b-tagging discriminants is the highest. Boosted top quark candidates are required to have $p_{\mathrm{T}} > 250$ GeV, exactly one constituent jet satisfying the 85% b-tagging working point plus at least one additional constituent jet not b-tagged. In case more than one boosted top quark candidate is identified, the one with the highest mass is selected.

Single lepton events containing at least one boosted Higgs boson candidate, at least one boosted top quark candidate and at least one additional jet b-tagged with the 85% WP enter the boosted signal region. On the contrary, events in the single-lepton channel not entering the boosted category need to have at least five jets with $p_{\mathrm{T}} > 25$ GeV and $|\eta| < 2.5$ and at least two of them have to be b-tagged using the 60% WP or three of them have to be b-tagged using the 70% WP in order to be further selected. Such events are classified as *resolved* single-lepton events.

Finally, events in the dilepton channel must have at least three jets with $p_{\mathrm{T}} > 25$ GeV and $|\eta| < 2.5$, of which at least two must be b-tagged with the 77% working point.

After the selection, events are subsequently classified into exclusive regions based on the number of jets and the number of b-jets tagged with the four different WPs, with the exception of events falling into the boosted region.

The four leading jets (in the SL) are used to finely categorize the selected events, e.g., an event that has three jets b-tagged at 60% and a fourth one at 77% will be classified in an analysis category different from an event with four jets all b-tagged at 60%. Sub-categories with a similar background composition are later grouped together. This

translates in having categories enriched in one of the relevant sample components: $t\bar{t}H$, $t\bar{t}+b\bar{b}$, $t\bar{t}+b$, $t\bar{t}+\geq 1c$ and $t\bar{t}+light$.

Analysis regions with an important contribution of $t\bar{t}H$ and $t\bar{t}+b\bar{b}$, relative to the other backgrounds, are considered as SR and further attempts are made to separate the $t\bar{t}H(b\bar{b})$ signal from the backgrounds. In contrast, the remaining control regions are used to derive constraints on backgrounds and systematic uncertainties in a likelihood fit.

In the SL channel a total of five signal regions are formed from events passing the resolved selection; three of them require at least six jets, while the remaining two require exactly five jets. They are referred to as $\mathrm{SR}_1^{\geq 6\mathrm{j}}$, $\mathrm{SR}_2^{\geq 6\mathrm{j}}$, $\mathrm{SR}_3^{\geq 6\mathrm{j}}$, $\mathrm{SR}_1^{5\mathrm{j}}$ and $\mathrm{SR}_2^{5\mathrm{j}}$. Events passing the boosted single-lepton selection form a sixth signal region, $\mathrm{SR}^{\mathrm{boosted}}$. The remaining events are then categorized into control regions enriched in $t\bar{t}+light$, $t\bar{t}+\geq 1c$ and $t\bar{t}+b$, for a total of six CR, three per jet multiplicity. The detailed definition of the signal and control regions for the resolved single-lepton channel is presented in Figure 5.15.

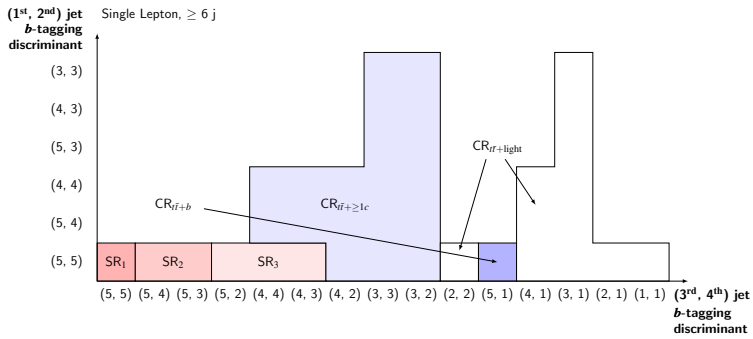
In a similar manner, three signal regions are defined in the dilepton channel, with different levels of purity for the $t\bar{t}H$ and $t\bar{t}+b\bar{b}$ components: $\mathrm{SR}_1^{\geq 4\mathrm{j}}$, $\mathrm{SR}_2^{\geq 4\mathrm{j}}$ and $\mathrm{SR}_3^{\geq 4\mathrm{j}}$. The remaining events are divided into four control regions: $\mathrm{CR}_{t\bar{t}+\mathrm{light}}^{\geq 4\mathrm{j}}$ and $\mathrm{CR}_{t\bar{t}+\mathrm{light}}^{3\mathrm{j}}$ enriched in $t\bar{t}+light$ and $\mathrm{CR}_{t\bar{t}+\geq 1c}^{\geq 4\mathrm{j}}$ and $\mathrm{CR}_{t\bar{t}+\geq 1b}^{3\mathrm{j}}$ enriched in $t\bar{t}+\geq 1c$ and $t\bar{t}+\geq 1b$.

Figure 5.16 shows the background composition for each of the analysis regions, whereas the $t\bar{t}H$ signal purity, $S/\sqrt{B}$, as well as the S/B ratio are shown in Figure 5.17.

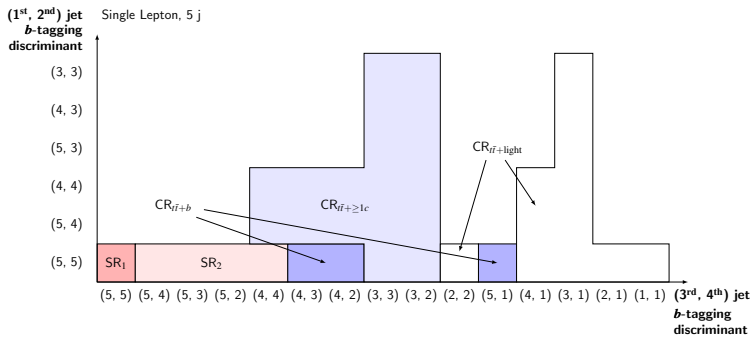
5.4.2 Final state reconstruction

As was the case for the ICHEP analysis, the final state reconstruction was achieved via the training of a BDT: the reconstruction BDT. In addition to that, two new techniques were implemented in order to enhance signal to background discrimination: a Likelihood Discriminant (LHD), that combines the signal and background probabilities of all possible combinations in each event; and a Matrix Element Method (MEM), that exploits the full matrix element calculation to separate the signal from the background.

Given that the three techniques make use of similar information from



(a)

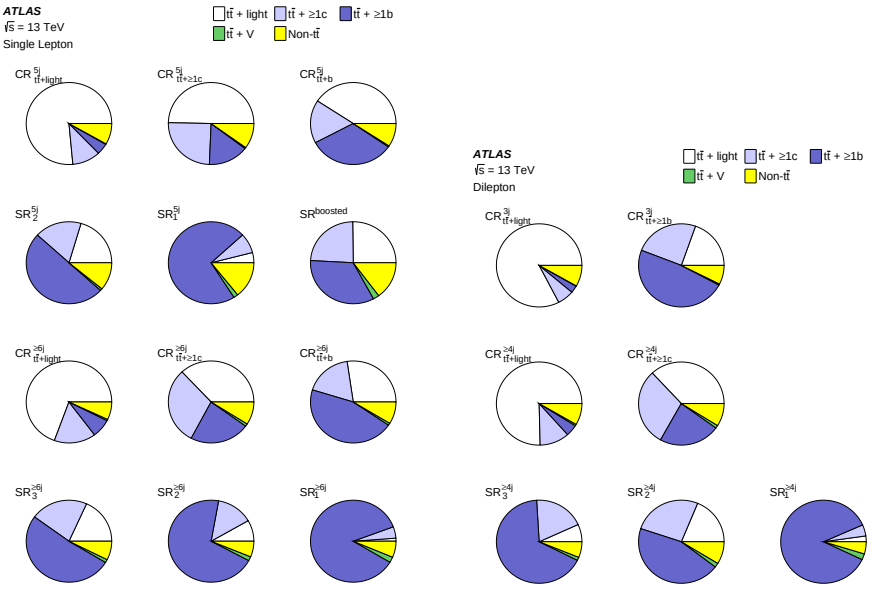


(b)

Figure 5.15: Definition of the six jet (a) and five jet (b) signal and control regions in the single-lepton resolved channel, as a function of the *b*-tagging discriminant. The vertical axis shows the values of the *b*-tagging discriminant for the first two jets, while the horizontal axis shows these values for the third and fourth jets. The jets are ordered according to their value of the *b*-tagging discriminant in descending order.

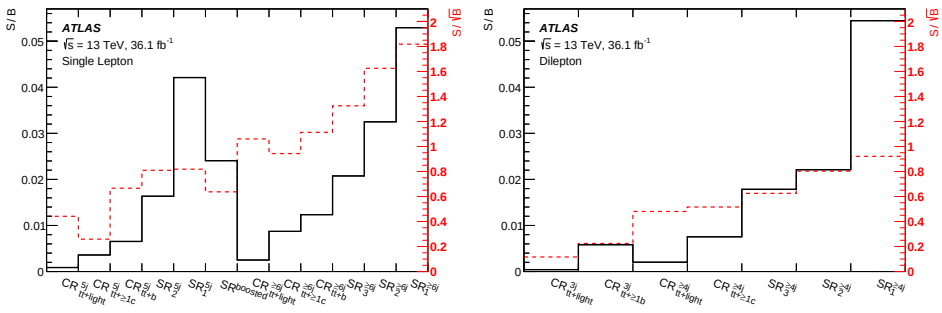
different perspectives and are based on different assumptions, the three methods show partial, not full, correlation among their outputs, indicating that not all the available information was exploited.

In the SR$^{\text{boosted}}$, on the other hand, there is no need to have advanced reconstruction techniques, as the Higgs boson and the top quark candidates are naturaly identified during the event categorization stage.



(a) **(b)**

Figure 5.16: Pie charts with the fractional contributions of the various backgrounds to the total background prediction in each of the analysis regions in the single-lepton channel (a) and in the dilepton channel (b).



(a) **(b)**

Figure 5.17: The ratios S/B (black solid line, referring to the vertical axis on the left) and $S/\sqrt{B}$ (red dashed line, referring to the vertical axis on the right) for each of the analysis categories in the single-lepton channel (a) and in the dilepton channel (b).

Reconstruction BDT

The reconstruction BDT is used to select the best jet-to-parton assignment in each event first and afterwards to build discriminating variables related to the Higgs boson and top quarks. It is employed in all dilepton and resolved single-lepton signal regions.

An improvement with respect to the implementation described in Section 5.3.2 comes from the possibility to use b-tagging information directly in the form of the bin in which the MV2c10 b-tagging discriminant of the jet falls. The leading four jets, sorted by their b-tagging weight bin, are considered as b-jets and the remaining jets are considered as light jets.

In order to simplify the analysis, the reconstruction BDT was trained inclusively in all the signal regions, with just a split based on the jet multiplicity. A comparison with dedicated trainings in all the different signal regions shows a similar performance.

Again, two versions of the recoBDT are used in the analysis: one that considers only variables related to the decay products of the $t\bar{t}$ system and one with additional variables related to the Higgs system. As representative values of the overall matching efficiencies, the Higgs boson is correctly reconstructed in 48% (32%) of the selected $t\bar{t}H$ events in the single-lepton channel $\mathrm{SR}_1^{\geq 6\mathrm{j}}$using the reconstruction BDT with (without) information about the Higgs boson kinematics included. For the dilepton channel, the corresponding reconstruction efficiencies are 49% (32%) in $\mathrm{SR}_1^{\geq 4\mathrm{j}}$.

Matrix element method

The Matrix Element Method (MEM) was already used in the Run1 search for the $t\bar{t}H(b\bar{b})$ process and its implementation follows very closely the one described in Ref. [174].

Likelihoods are constructed to express the degree to which the event is consistent with a specific physics process: the $t\bar{t}H(b\bar{b})$ signal and the $t\bar{t}+b\bar{b}$ background. The discriminating variable is then defined as the difference between the logarithms of the signal and background likelihoods: $\mathrm{MEM}_{D1} = \log_{10}(L_S) - \log_{10}(L_B)$.

Each likelihood is a sum over multiple jet-to-parton assignments, but

in order to reduce computation time, only initial states induced by gluons are considered. The likelihoods for both hypotheses are then computed using matrix element calculations at parton level and then transfer functions are used to map the reconstructed detector quantities to the corresponding parton level quantities. In this way it is possible to link theoretical calculations to observed quantities, making the most complete use of the kinematic information of a given event.

Given that MEM consumes a significant amount of computation time, it is implemented only in the most sensitive single-lepton signal region, $\mathrm{SR}_1^{\geq 6\mathrm{j}}$. Furthermore, b-tagging information is used to reduce the number of jet-to-parton assignments considered in the calculation.

The MEM_{D1} discriminant is then included as another input to the classificationBDT.

Likelihood discriminant

In the signal regions of the resolved single-lepton channel, a likelihood discriminant is employed to discriminate between the signal and background hypothesis. The discriminant is defined as:

$$\mathrm{LHD} = \frac{p^{\mathrm{sig}}}{p^{\mathrm{sig}} + p^{\mathrm{bkg}}} \tag{5.6}$$

where p^{sig} and p^{bkg} represent the probability density functions of a given event under the signal and background hypothesis respectively.

These probabilities are obtained as the product of one-dimensional MC-based probability density functions, built from various kinematic distributions, such as invariant masses and angular variables, averaged among all possible jet-to-parton matching assignments. Each parton-to-jet assignment is weighted using the b-tagging information to give more importance to permutations whose parton matching is more consistent with the correct flavour of the parton candidates.

Two background hypotheses are considered, corresponding to the production of $t\bar{t}+b$ and $t\bar{t}+\geq 2$ b-jets. The final value of the discriminant is an average of the LHD for both hypotheses, weighted by their relative fractions in simulated events, which are approximately 20% and 80% respectively.

Furthermore, in a significant fraction of $t\bar{t}H$ and $t\bar{t}$ simulated events with at least six selected jets, only one jet stemming from the hadronically decaying W boson is selected, as the other falls outside the acceptance region; therefore, an additional hypothesis, for both the signal and the background, is considered to account for this missing jet topology.

One advantage of the LHD method over the reconstruction BDT is that it takes advantage of all possible combinations in the event, but on the other hand it does not fully account for correlations between variables in one combination, as it uses a simple product of one-dimensional probability density functions.

The likelihood discriminant is then added as an input variable to the classification BDT.

5.4.3 BDT to classify the events

After the event categorization, classification BDTs were trained in the SRs in a similar way as it was done for the ICHEP analysis.

Several variables are combined into each classification BDT, each of them exploiting the different kinematic aspects of signal and background events. Additionally, b-tagging information and the outputs of the intermediate multivariate discriminants (MEM_{D1}, recoBDT and LHD) are used for this round, which represent the most powerful variables in the classification BDT.

For the boosted signal region, kinematic variables are built from the properties of the large-R jets and their jet constituents.

The full list of input variables in each of the signal regions is presented in Appendix E.

5.5 Intermezzo: statistical analysis

In order to extract the maximum information from the collected data, a profile likelihood fit is performed, given the ability of this approach to reduce the impact of systematic uncertainties on the final results. The material presented is taken from Refs. [175–177].

5.5.1 The profile likelihood method

To summarize the outcome of a search, the level of agreement of the observed data with a given hypothesis needs to be computed, therefore the first steps towards a statistical analysis of the collected data is to define a statistical model of the data itself, which contains all the understanding of the underlying physics.

The hypotheses that are tested are the background-only hypothesis, where no $t\bar{t}H$ process is present, against the signal-plus-background hypothesis, which coincides with the presence of the associated production precisely described by the SM. The two are defined by using a parameter μ, called *signal strength*, defined as the ratio of the measured cross-section over the expected one from the Standard Model:

$$\mu = \frac{\sigma_{\text{meas}}}{\sigma_{\text{SM}}} \tag{5.7}$$

so that $\mu = 0$ identifies the background-only and $\mu = 1$ and the signal-plus-background hypothesis.

All the physics knowledge enters in the definition of a likelihood function, $\mathscr{L}(data, \mu, \theta)$, that depends on the observed distribution of the data, the *parameter of interest* μ and contains term used to encode the effects of the systematic uncertainties:

$$\mathscr{L} = \prod_{i}^{reg} \prod_{j}^{bins} \text{Pois}(n_{ij}|\mu s_{ij} + b_{ij}) \prod_{k} \text{Gaus}(\theta_k|0,1) \tag{5.8}$$

The first term of the likelihood is constructed as a product of Poisson probability terms over all bins of the chosen distribution of each region: n_{ij} is the number of data events, s_{ij} and b_{ij} represent the expected number of signal and background events in the j-th bin of the input distribution in the region i. The product of Gaussian distributions contains the modelling of the systematics, described with the inclusion of a set of continuous parameters, θ_k, that parametrize the effect of each systematic on the signal and backgrounds templates distributions: varying θ allows to modify both the shape and normalization of the templates. The only parameter of interest is μ and all other adjustable parameters needed to specify the model are called *nuisance parameters* (NP).

Each of the various systematic uncertainties comes from a dedicated auxiliary measurement, which gives the central value and the uncertainty of the systematic effect under consideration, e.g., the jet energy scale or scale factors associated with b-tagging calibrations. Without loss of generality, it's common practice to parametrize these measurements so that the nominal value corresponds to $\theta = 0$ and values of ± 1 correspond to the $\pm 1\sigma$ variations.

Furthermore, the statistical uncertainty on the predicted MC templates is included in the likelihood definition by introducing additional nuisance parameters, one for each bin considered. This allows the total prediction to fluctuate within the statistical uncertainty.

The best estimate for μ, as well as for the θ parameters, is obtained by maximising the likelihood function in Eq. (5.8) or equivalently minimising the negative logarithm of the likelihood. The best-fit values are indicated with $\hat{\mu}$ and $\hat{\theta}$ in the following, while the conditional maximum likelihood estimate, $\hat{\hat{\theta}}(\mu)$, is defined as the value that maximizes the likelihood for a fixed value of μ. Note that $\hat{\mu}$ is allowed to be negative.

The procedure of choosing a specific value of the NPs for a given μ is often referred to as *profiling*. The profile likelihood ratio is defined as:

$$\lambda(\mu) = \frac{\mathscr{L}(\mu, \hat{\hat{\theta}}(\mu))}{\mathscr{L}(\hat{\mu}, \hat{\theta})} \tag{5.9}$$

This approach allows the data under study to potentially improve the initial knowledge of systematic uncertainties obtained via dedicated measurements. If the selected data is not sensitive to a given source of systematic uncertainty, the constraint term in the likelihood ensures that the nuisance parameter stays at 0 and its error corresponds exactly to the input uncertainty. On the other hand, if the effect of one or more systematic uncertainties is not supported by the data, the fit procedure could shift (pull) the central value of a nuisance parameter to achieve a better data/MC description or produce a reduction (constraint) of the error of the nuisance parameter with respect to its initial value.

Furthermore, during the minimization process, correlations among NP can arise spontaneously depending of their effect.

The main reason for using the profile likelihood ratio is that its asymptotic distribution, i.e. with a large number of events, does not depend on the nuisance parameters.

5.5.2 Asymptotic limit and expected results

For a sufficiently large data sample, approximations exists for the profile likelihood ratio, based on the results of Wilks and Wald [178, 179]. It is asymptotically related to the χ^2 distribution for one degree of freedom per parameter of interest. In particular, if the data is distributed according to an underlying true signal strength parameter μ', then the distribution of the test statistic can be expressed as:

$$-2\ln\lambda(\mu) = \frac{(\mu-\hat{\mu})^2}{\sigma^2} + \mathcal{O}(1/\sqrt{N}) \tag{5.10}$$

where the fitted signal strength $\hat{\mu}$ follows a Gaussian distribution with mean μ' and standard deviation σ and N is the data sample size[5].

The variance σ^2 can be estimated from an artificial dataset, referred to as *Asimov dataset*[6] [175] that is constructed by generating pseudo-data that matches exactly the number of background and signal events, for a given value of μ, expected in every bin of the input distributions.

In the Asimov dataset all the statistical fluctuations are suppressed, thus when this dataset is used to evaluate the estimators for all parameters, the true values of the parameters are obtained.

Furthermore, the Asimov dataset can be used to report not only the expected significance, but also the range of values in which the significance is expected to vary, given that the collected dataset will necessarily have statistical fluctuations.

5.5.3 Significance, signal discovery and upper limits

To establish a discovery, the hypothesis test is formulated in terms of rejecting the null hypothesis ($\mu = 0$, i.e. there is no Higgs boson signal

[5] In practice the approximations are found to provide an accurate description even for fairly small data samples, such as 20 or so.

[6] The name Asimov is inspired by the short story *Franchise*, by Isaac Asimov. In it, elections are held by selecting the single most representative voter in place of the entire electorate.

present): roughly speaking, claiming a discovery is equivalent of stating that the observed data is incompatible with the null hypothesis.

A robust way of defining a test statistic in case $\mu \geq 0$ is the following:

$$q_0 = \begin{cases} -2\ln\lambda(0) & \text{if } \hat{\mu} \geq 0 \\ 0 & \text{if } \hat{\mu} < 0 \end{cases} \tag{5.11}$$

If data fluctuates so that the best-fit values is $\hat{\mu} < 0$, the corresponding value is $q_0 = 0$. As $\hat{\mu}$ and the event yields increase above the expected background, q_0 increases accordingly, corresponding to a higher level of incompatibility between data and the $\mu = 0$ hypothesis.

Eventually, to a given dataset corresponds an observed value $q_{0,\text{obs}}$. The level of disagreement between the data and the null hypothesis can thus be quantified by the computation of the p_0-value:

$$p_0 = \int_{q_{0,\text{obs}}}^{\infty} f(q_0|\mu = 0)\,\mathrm{d}q_0. \tag{5.12}$$

where $f(q_0|\mu = 0)$ denotes the probability density function of the test statistics q_0 under the null hypothesis. The p-value is therefore a measure of the probability of observing a dataset as signal-like or more as the actual observed dataset, under the assumption that there is no signal. Small p_0-values are interpreted as evidence against $\mu = 0$.

It is common practice to define the *significance* that corresponds to a given p_0-value as the number of standard deviations, Z, at which a Gaussian distributed variable with zero mean and variance equal to 1 would give a *one-sided* tail area equal to the measured p_0-value.

In particle physics the value of $Z = 5$ is used to reject the background-only hypothesis and claim a discovery[7], which corresponds to a p_0-value of $p_0 = 2.87 \times 10^{-7}$.

On the other hand, with the profile likelihood ratio formalism it is also possible to reject the signal-plus-background hypothesis for some values of μ different from zero. This procedure will result in a range of values excluded by the data at a certain confidence level (CL), typically CL = 95% is used, which corresponds to $Z = 1.64$.

[7] The interested reader will find in Ref. [180] a nice discussion on the origin of this particularly high threshold.

Roughly speaking, it is the opposite of the discovery case. In practice, this is done by finding the p-value such that p_μ = 1 - CL and solving for μ. The p_μ-value is defined in an analogous way to the discovery case:

$$p_\mu = \int_{q_{\mu,\text{obs}}}^{\infty} f(q_\mu|\mu)\,\mathrm{d}q_\mu. \tag{5.13}$$

where $f(q_\mu|\mu)$ is the probability density function of the test statistic q_μ, defined as:

$$q_\mu = \begin{cases} -2\ln\lambda(\mu) & \text{if } \hat{\mu} \leq \mu \\ 0 & \text{if } \hat{\mu} > \mu \end{cases} \tag{5.14}$$

It is important to note that it is not a simple generalization of the Eq. (5.11), but it has its own definition; as a matter of fact $q_0 = 0$ if the data fluctuate downward, but $q_\mu = 0$ if the data fluctuate upward ($\hat{\mu} > \mu$).

An interesting case is when the hypothesis $\mu = 1$ can be rejected: the corresponding signal hypothesis is considered as excluded.

In case the distributions of the test statistic q_μ, for the null and the alternative hypothesis, are very close to each other, i.e. the experiment has low sensitivity, the p-value can reject a model even if there is not enough sensitivity due to downward fluctuations in the observed data. In order to overcome this problem, a modified method was introduced: the CLs method [181], which uses the variable:

$$CL_s = \frac{CL_{s+b}}{CL_b} = \frac{p_\mu(\mu = 1)}{1 - p_0} \tag{5.15}$$

where p_μ is the p-value computed using Eq. (5.13).

In LHC searches, the variable CL_s is used to set upper limits instead of p_μ. If $CL_s < 0.05$, the signal-plus-background hypothesis is excluded at 95% confidence level.

5.6 Experimental results

In this section the fit model and the experimental results will be presented.

5.6.1 The fit model

As said multiple times throughout this work, a profile likelihood ratio fit is performed to data.

The chosen distributions for the signal regions of both the DIL and SL (resolved and boosted) channels are the classification BDT outputs, with the binning of the input distributions optimized in order to maximize the analysis sensitivity. Among the various control regions, only the $\mathrm{CR}^{\geq 6\mathrm{j}}_{t\bar{t}+\geq 1c}$ and $\mathrm{CR}^{5\mathrm{j}}_{t\bar{t}+\geq 1c}$ employ as discriminant the H_{T} variable, defined as the scalar sum of the p_{T} of all the jets, whereas all the other ones enter simply as a one-bin distribution, i.e. only the total number of the events is used as input. The decision of not using the H_{T} distribution in the other CR is because studies on the blinded dataset showed the presence of pulls and constraints of some NP beyond what is considered to be acceptable.

Only one signal strength parameter common to both channels is used in the fit. The $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ backgrounds, the two most important ones, are both assigned a free-floating normalization factor, κ_{ttb} and κ_{ttc}, which are only constrained by the fit to data and are used by the fit to absorb normalization mismodelling of the corresponding backgrounds. This is necessary due to the discrepancy between the observed data yields and the MC prediction especially in the regions where the $t\bar{t}$+HF component is predominant, as it is known from previous studies [162, 163] and is visible in Figures 5.18a and 5.18c. In this way the signal extraction will not be biased by a general underestimation of the predicted backgrounds.

The scheme used to incorporate the various sources of systematic uncertainties in the likelihood definition is of equal importance. The origins of these uncertainties include both experimental and theoretical sources, such as the reconstruction and identification of leptons and jets or the modelling of the signal and background processes.

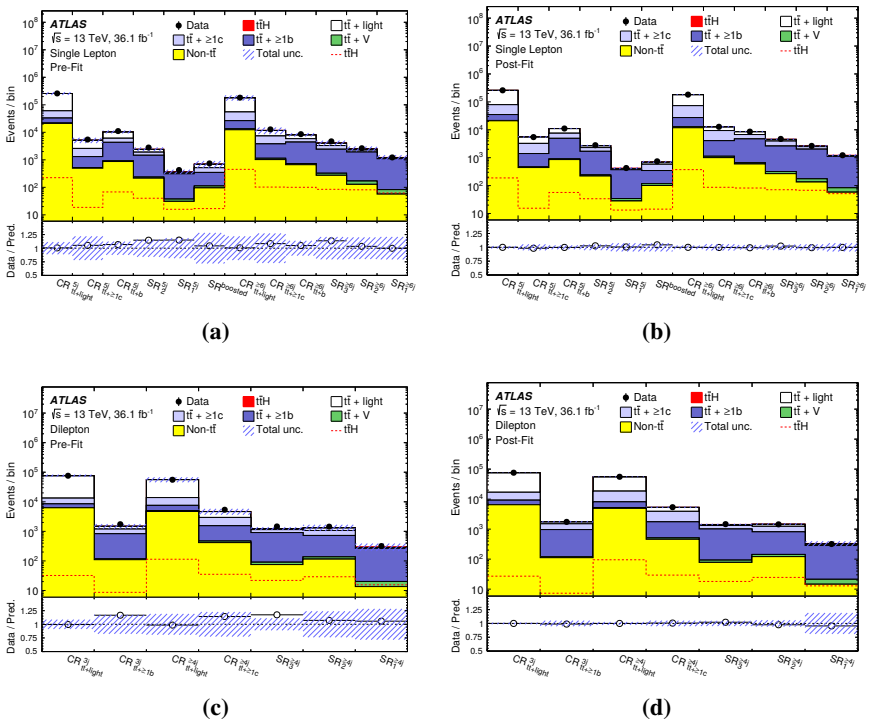


Figure 5.18: Comparison of predicted and observed event yields in each of the control and signal regions, in the semilepton (top) and dilepton (bottom) channels before (left) and after (right) the fit to the data.

They can affect both the normalization and shape of the various samples considered in the search, with the exception of the luminosity and cross-section uncertainties, which affect only the normalization. In spite of that, the normalization uncertainties can and do modify the relative fractions of the different samples, which leads to a change in the shape of the final discriminant distribution under consideration.

Individual sources of systematic uncertainty are considered uncorrelated, whilst each source has a correlated effect across the boosted, single-lepton and dilepton channels, their regions and their samples. Furthermore, most of the experiment uncertainties are decomposed in several orthogonal components.

Lastly, if a systematic source has an effect less than 0.5% in changing the normalization or the shape of a sample in one region, it is removed for that specific sample and region. This procedure is called *pruning* and is employed in order to simplify the model and speed up the evaluation time. Detailed comparisons have been carried out and they showed that no difference in the results occurs due to this procedure.

Experimental uncertainties

The uncertainties related to the object reconstruction have been described in Chapter 3, thus only a brief description will be reported here.

The uncertainty on the total integrated luminosity for the combined 2015+2016 dataset is 2.1%. It is derived following a similar methodology to the one detailed in Ref. [155], from a calibration of the luminosity scale using x–y beam-separation scans performed in August 2015 and May 2016.

A variation in the pileup reweighting of MC events is included to cover the uncertainty in the ratio of the predicted and measured inelastic cross-sections in the fiducial volume defined by $M_X > 13$ GeV, where M_X is the mass of the hadronic system [156].

The jet energy scale uncertainty is derived by combining several information and is factorized into eight independent components. Further sources are considered, which are related to the jet flavour composition, pileup corrections and η-intercalibration, high-p_T jets, jet energy resolution and the efficiencies of the pileup suppression cut, as described in Section 3.3, for a total of 21 independent jet-related systematic uncertainties.

Calibration correction factors of the efficiencies of the flavour tagging algorithm to correctly identify the three flavour components are used in the analysis and the uncertainties on the correction factors are considered as well. The pseudo-continuous b-tagging introduces complications due to the use of several working points simultaneously. The b-tagging efficiencies and mis-tag rate are first measured for the four working points separately and later combined in the calibration of the whole MV2c10 discriminant distribution, with care in considering the correlation among the various MV2c10 bins. The uncertainties are later factorized into 30 independent components associated with the b-jet

tagging efficiency, 15 component for the c-jets and 80 for light-jets.

Lepton identification, isolation and reconstruction efficiency, as well as trigger efficiencies and lepton momentum scale and resolution, have systematic uncertainties associated with them. These are measured in data, as explained in Section 3.2, and account for a total of 24 independent sources.

Lastly, uncertainties in the scale and resolution of the missing energy soft term are considered, for a total of three additional sources of systematic uncertainty.

Signal and $t\bar{t}$ modelling uncertainties

Two independent sources of uncertainties are associated with the $t\bar{t}H$ cross-section: the QCD scale uncertainty and the PDF+α_S one [166–171], for an uncertainty of $^{+5.8\%}_{-9.2\%}$ (scale) $\pm 3.6\%$ (PDF). In addition, uncertainties on the theoretical Higgs boson branching fractions are considered, which amount to 2.2% for the $b\bar{b}$ decay mode [166]. The last uncertainty on the $t\bar{t}H$ signal is associated with the choice of the parton shower and hadronization model, derived by comparing the nominal prediction to the one obtained with events generated by MADGRAPH5_aMC@NLO interfaced to Herwig++.

A 6% normalization uncertainty is considered for the inclusive $t\bar{t}$ production cross-section at NNLO+NNLL [133], which includes the effects from varying the factorization and renormalization scales, the PDF, α_S and the top quark mass. This is the only systematic uncertainty that is correlated among the three $t\bar{t} + \geq 1b$, $t\bar{t} + \geq 1c$ and $t\bar{t} + light$ categories.

The other $t\bar{t}$ modelling uncertainties either affect only one of the three $t\bar{t}$ + jets components or are considered uncorrelated among them, given that the $t\bar{t} + light$ profits from precise measurements in data, while this is not the case for the other two components. In addition, the mass difference between the b- and c-quark contributes to a difference between the two processes and the flavour scheme used for the PDF: 4FS vs 5FS. The normalizations of $t\bar{t} + \geq 1b$ and $t\bar{t} + \geq 1c$ are allowed to float freely in the fit.

A comparison between the nominal POWHEG+PYTHIA8 sample and

the SHERPA5F one provides the uncertainty associated with the choice of the $t\bar{t}$ inclusive generator for the simulation of the hard scatter, even if it is obtained by actually varying both the generator and the parton shower and hadronization model. In order to have a fair comparison, the SHERPA5F sample, along with all the other alternative samples, underwent the same reweighting procedure exposed in Section 5.2.2, i.e. the subcategories of the $t\bar{t} + \geq 1b$ sample are scaled to match the predictions of SHERPA4F. Furthermore, the alternative samples are reweighted in such a way that their $t\bar{t} + \geq 1b$ and $t\bar{t} + \geq 1c$ fractions match the one in the nominal sample.

Similarly to what was done for the calibration of Jet Vertex Charge, the parton shower and hadronization model uncertainty is derived by comparing the nominal POWHEG+PYTHIA8 with the predictions from POWHEG interfaced with Herwig7, whereas the uncertainty in the modelling of initial and final state radiation is assessed with two alternative POWHEG+PYTHIA8 samples with "up" and "down" variations. As an example, Figure 5.19 shows the effect of the generator uncertainty and the uncertainty associated to the choice of parton shower model on the $t\bar{t} + \geq 1b$ templates in the $\mathrm{SR}_1^{\geq 6\mathrm{j}}$.

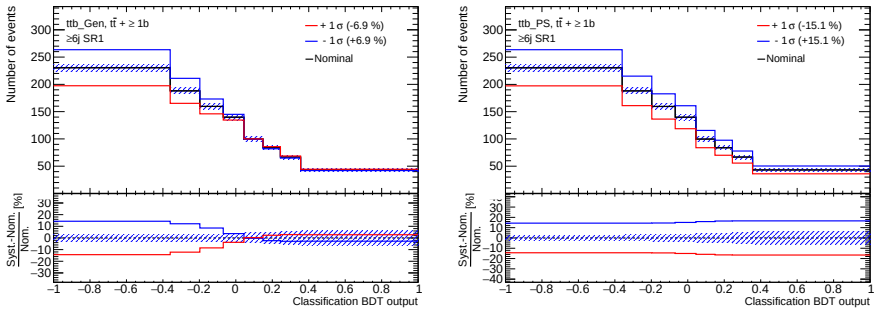
Given the difficulties of describing the $t\bar{t} + \geq 1c$ background from a theoretical point of view and the poor experimental guidance, an ad hoc uncertainty is applied to this background. The systematic is derived by comparing the nominal sample to an NLO sample of $t\bar{t}$ $+c\bar{c}$ in the matrix element, including massive c-quarks (effectively a 3F scheme), produced with MADGRAPH5_aMC@NLO interfaced to Herwig++, as described in Ref. [182]. This uncertainty is related to the choice between the $t\bar{t}$ $+c\bar{c}$ ME calculation and the prediction from an inclusive $t\bar{t}$ sample, where the c-jets are mainly produced in the parton shower process.

Due to the importance of the $t\bar{t} + \geq 1b$ background, several additional specific uncertainties have been considered. Different descriptions of this process can be obtained either with a dedicated NLO calculation of the $t\bar{t} + \geq 1b$ ME in a 4F scheme generator or with the nominal POWHEG+PYTHIA8 inclusive $t\bar{t}$ (5F) sample. A comparison of the two samples is used to derive the corresponding systematic uncertainty.

Uncertainties modifying the relative fraction of the $t\bar{t} + b$, $t\bar{t} + b\bar{b}$,

$t\bar{t}+B$ and $t\bar{t}+\geq 3b$ subcomponents of the SHERPA4F sample are taken into account. They are derived by varying parameters internal to the SHERPA generator, as well as by considering two alternative PDF sets; these uncertainties are the ones contributing to the uncertainty band shown in Figure 5.3 for the SHERPA4F prediction. Additionally, a 50% normalization uncertainty is assigned to the $t\bar{t}+\geq 3b$ process, given that the discrepancy between the 4F and the 5F prediction is not covered by the aforementioned systematics.

Lastly, a 50% normalization uncertainty is considered on the MPI contribution, based on studies of different underlying event sets of tuned parameters, as the fraction of this subcategory is not fixed in the alternative samples.



(a) **(b)**

Figure 5.19: Effect of the generator (a) and parton shower and hadronization (b) systematic uncertainties on the $t\bar{t}+\geq 1b$ background template in the $\mathrm{SR}_1^{\geq 6\mathrm{j}}$ signal region.

Modelling of the other backgrounds

Among the non-$t\bar{t}$ processes, the W/Z+jets, single top and fakes backgrounds are the most important ones, even though they represent a minor fraction of the total background.

Two uncertainties are assigned to the W+jets cross-section: an overall 40% normalization and an additional 30% normalization uncertainty only for events with heavy-flavour jets, which is uncorrelated between

events with two and more than two of such jets. The Z+jets has a 35% uncertainty applied uncorrelated for events with different jet multiplicities. These uncertainties are based on variations of the factorization and renormalization scales, as well as matching parameters in the SHERPA simulation.

The three cross-sections for the single-top production modes, namely the s-channel, the t-channel and the Wt-channel, get a $^{+5\%}_{-4\%}$ uncertainty each [141–143]. An uncertainty in the amount of interference between Wt and $t\bar{t}$ production at NLO [145] is assessed by comparing the default "diagram removal" scheme to the alternative "diagram subtraction" scheme. The last two uncertainties on the single-top production are related to the choice of parton shower and hadronization model on one side and the amount of radiation on the other, for both the Wt and t-channels, for a total of four systematics. They are evaluated by comparing the nominal samples with ad hoc samples that use alternative settings in full analogy of what is done for the $t\bar{t}$ sample.

A 50% normalization uncertainty is assumed for the diboson background, which includes both the uncertainty on the inclusive cross-section and additional jet production [150].

A 50% normalization uncertainty is assigned to the overall prediction of the fakes background, uncorrelated between the electron+jets and muon+jets channels, uncorrelated between with regions with 5 and 6 jets and between the resolved and boosted channels. In the dilepton channel, to this background a 25% uncertainty is assigned, correlated across lepton flavours and all analysis regions.

The $t\bar{t}+W/Z$ NLO cross-section prediction uncertainty is 15% [183]. Additional modelling uncertainties related to the choice of the matrix element generator, parton shower and hadronization are evaluated, as usual, by comparing the nominal $t\bar{t}V$ samples to alternative one generated with SHERPA. A generic 50% normalization uncertainty is assigned to the $t\bar{t}t\bar{t}$ background. The backgrounds from tZ, $t\bar{t}WW$, $tHjb$ and WtH are each assigned two normalization uncertainties related to PDF and scale variations, while to tWZ is assigned one cross-section uncertainty that accounts for both the scale and PDF effects.

5.6.2 Results

The fit in all the single-lepton (both resolved and boosted) and dilepton regions yields a best-fit value for the signal strength of:

$$\mu = 0.84 \pm 0.29\ (\text{stat.})\ ^{+0.57}_{-0.54}\ (\text{syst.}) = 0.84^{+0.64}_{-0.61} \tag{5.16}$$

and the expected uncertainty of the signal strength is identical to the measured one.

The observed signal strengths for the individual channels and their comparison with the combined one is summarized in Figure 5.20. The two signal strengths are obtained with a combined fit in which the two μ's are independent of each other, but the nuisance parameters are correlated as in the single-μ fit. On the other hand, fitting the two channels separately yields the observed signal strengths to be $\mu = 0.11^{+1.36}_{-1.41}$ and $\mu = 0.67^{+0.71}_{-0.69}$ for the dilepton and single-lepton respectively. The fact that both values are lower than the combined one is due to the large correlations in the systematic uncertainties of the background prediction between the two channels.

The H_{T} distributions in the $\mathrm{CR}^{5\mathrm{j}}_{t\bar{t}+\geq 1c}$ and $\mathrm{CR}^{\geq 6\mathrm{j}}_{t\bar{t}+\geq 1c}$ both pre- and post-fit are shown in Figure 5.21 and the classification BDT output distributions are presented in Figures 5.22 to 5.24.

All these distributions are reasonably well modelled pre-fit within the assigned uncertainties and the fit is able to improve the level of agreement by adjusting the nuisance parameters; this is particularly true for the best-fit values of $\kappa_{\mathrm{ttb}} = 1.24 \pm 0.10$ and $\kappa_{\mathrm{ttc}} = 1.63 \pm 0.23$. The post-fit uncertainty is also significantly reduced, as a result of the constraints on the nuisance parameters, as well as the correlations generated by the likelihood fit.

The total error on the signal strength is dominated by the systematic component, whose main contribution comes from the uncertainties in the modelling of the $t\bar{t} + \geq 1b$ background, followed by the limited number of events in the simulated samples ("background-model stat. unc."), the flavour tagging uncertainties, the jet energy scale and resolution and the modelling of signal process. The total stat. uncertainty includes the uncertainties associated with the $t\bar{t}$+HF normalizations and it is obtained by redoing the fit to data after all the NP are fixed to their

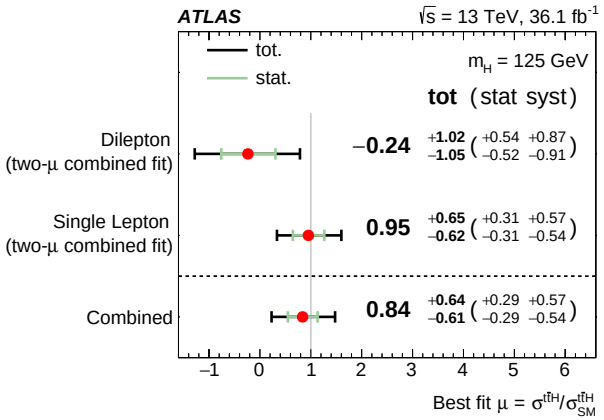


Figure 5.20: Summary of the signal-strength measurements in the individual channels and for the combination. All the numbers are obtained from a simultaneous fit in the two channels, but the measurements in the two channels separately are obtained keeping the signal strengths uncorrelated, while all the nuisance parameters are kept correlated across channels.

post-fit values. It is not simply the sum in quadrature of the κ_{ttb}, κ_{ttc} and the intrinsic statistical uncertainty due to the presence of correlations among them. Table 5.4 summarizes the contributions of the different uncertainties in the combined fit, grouped based on their source.

Figure 5.25, on the other hand, shows only the top 20 individual nuisance parameters with the largest impact on the total uncertainty on the signal strength, ranked in descending order. For each of the NP, the best-fit value and the post-fit uncertainty are shown. Performing the fit excluding the systematic uncertainties not present in this figure reduces the total uncertainty on the measured μ by 5%.

The black points show the pulls of the NP relative to their nominal values, θ_0, and their relative post-fit errors, $\Delta\hat{\theta}/\Delta\theta$. They both refer to the scale at the bottom of the plot. The empty (filled) blue rectangles correspond to the pre-fit (post-fit) impact on μ, both referring to the upper scale. The impact of each NP, $\Delta\mu$, is computed by comparing the nominal best-fit value of μ with the result of the fit when fixing

Table 5.4: Breakdown of the contributions to the uncertainties in μ. The "background-model stat. unc." refers to the statistical uncertainties in the MC events and in the data-driven determination of fake leptons background component in the single-lepton channel. The total uncertainty is different from the sum in quadrature of the different components due to correlations between nuisance parameters built by the fit.

Uncertainty source	$\Delta\mu$	
$t\bar{t} + \geq 1b$ modelling	$+0.46$	-0.46
Background-model stat. unc.	$+0.29$	-0.31
b-tagging efficiency and mis-tag rates	$+0.16$	-0.16
Jet energy scale and resolution	$+0.14$	-0.14
$t\bar{t}H$ modelling	$+0.22$	-0.05
$t\bar{t} + \geq 1c$ modelling	$+0.09$	-0.11
JVT, pileup modelling	$+0.03$	-0.05
Other background modelling	$+0.08$	-0.08
$t\bar{t} + light$ modelling	$+0.06$	-0.03
Luminosity	$+0.03$	-0.02
Light lepton (e, μ) id., isolation, trigger	$+0.03$	-0.04
Total systematic uncertainty	$+0.57$	-0.54
$t\bar{t} + \geq 1b$ normalization	$+0.09$	-0.10
$t\bar{t} + \geq 1c$ normalization	$+0.02$	-0.03
Intrinsic statistical uncertainty	$+0.21$	-0.20
Total statistical uncertainty	$+0.29$	-0.29
Total uncertainty	$+0.64$	-0.61

the considered NP to its best-fit value, $\hat{\theta}$, shifted by its pre-fit (post-fit) uncertainties $\pm\Delta\theta$ ($\pm\Delta\hat{\theta}$).

The uncertainty with the largest impact on the signal strength is the one coming from the comparison between SHERPA5F and the nominal prediction coming from POWHEG+PYTHIA8 for the $t\bar{t} + \geq 1b$ process. Three other uncertainties related to the modelling of the $t\bar{t} + \geq 1b$ background are immediately following. Concerning the theoretical uncertainties, the $t\bar{t}H$ signal modelling and the modelling of the $t\bar{t} + \geq 1c$

and $t\bar{t} + light$ backgrounds appear as well, while among the experimental systematics the most important ones are related to b-tagging and the jet energy scale and resolution; however, their contributions are significantly smaller than the ones from the $t\bar{t} + \geq 1b$ modelling.

Some of the nuisance parameters in Figure 5.25 are shifted by the fit from their nominal values. In order to understand the origin of such shifts, the corresponding NP is temporarily decorrelated across the analysis regions and samples and the fit is repeated. Typically, only one sample or region is responsible for the shift.

These shifts are used by the fit mostly to correct the $t\bar{t}$ background predictions to match the observed data in various regions. Similar shifts are seen when a background-only fit is performed after dropping the bins with the highest signal contributions, supporting the idea that the origin of the shifts lies in the background modelling.

An excess of events over the expected SM background is found with an observed (expected) significance of 1.4 (1.6) standard deviations. A signal strength larger than 2.0 is excluded at the 95% confidence level, as shown in Figure 5.26. The expected significance and exclusion limits are calculated using the background estimate after the fit to the data. The limits for the two individual channels are derived consistently with Figure 5.20, keeping the nuisance parameters correlated between both channels, but with independent signal strengths.

Figure 5.27 shows the event yield in data compared to the post-fit prediction for all events entering the analysis selection, grouped and ordered by the signal-to-background ratio of the corresponding final-discriminant bins. The predictions are shown for both the fit with the background-only hypothesis and with the signal-plus-background hypothesis, where the signal is scaled to either the measured μ or the value of the upper limit on μ.

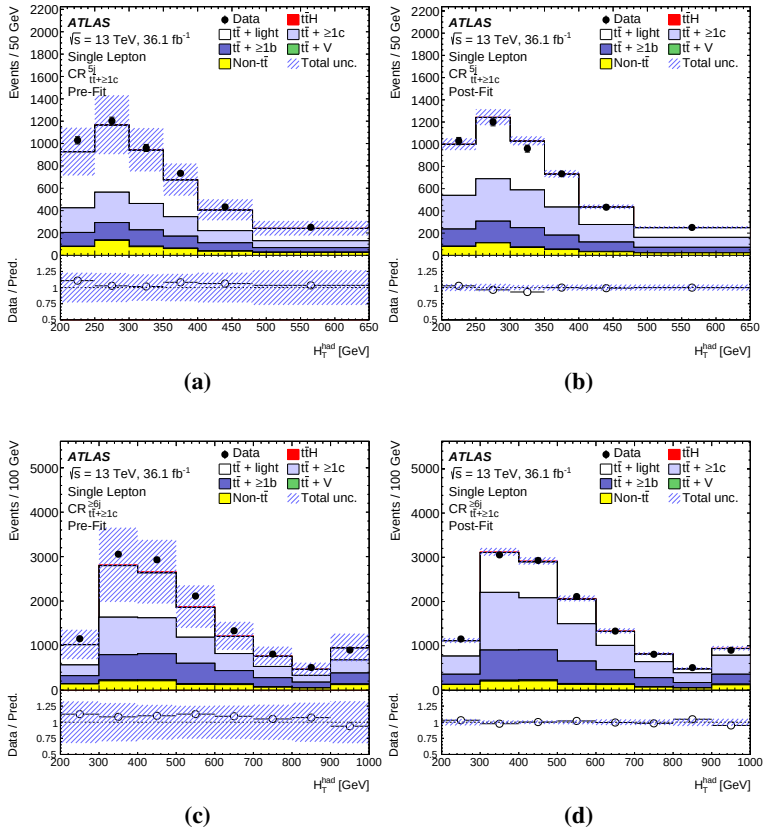


Figure 5.21: Comparison between data and prediction for the $H_{\mathrm{T}}^{\mathrm{had}}$ distributions in the single-lepton $t\bar{t} + \geq 1c$ enriched control regions the combined dilepton and single-lepton fit to data. The $t\bar{t}H$ signal prediction is shown stacked at the top of the background prediction, normalized to the SM cross section before the fit and to the fitted μ after the fit. The pre-fit plots do not include an uncertainty for the $t\bar{t} + \geq 1b$ or $t\bar{t} + \geq 1c$ normalization.

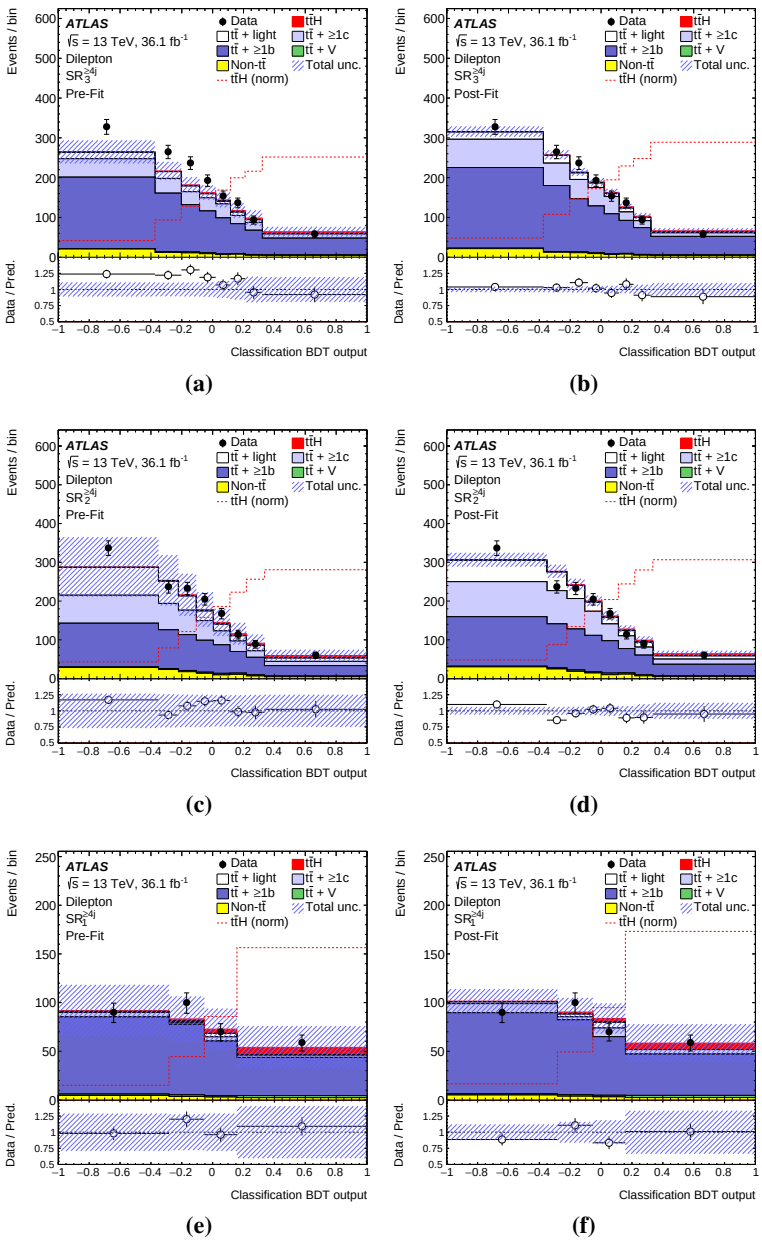


Figure 5.22: Comparison between data and prediction for the classification BDT discriminant in the dilepton signal regions pre- (left) and post- (right) the combined dilepton and single-lepton fit to data. The $t\bar{t}H$ contribution is normalized to the SM cross section before the fit and to the fitted μ after the fit. The pre-fit plots do not include an uncertainty for the $t\bar{t}+\geq 1b$ or $t\bar{t}+\geq 1c$ normalization.

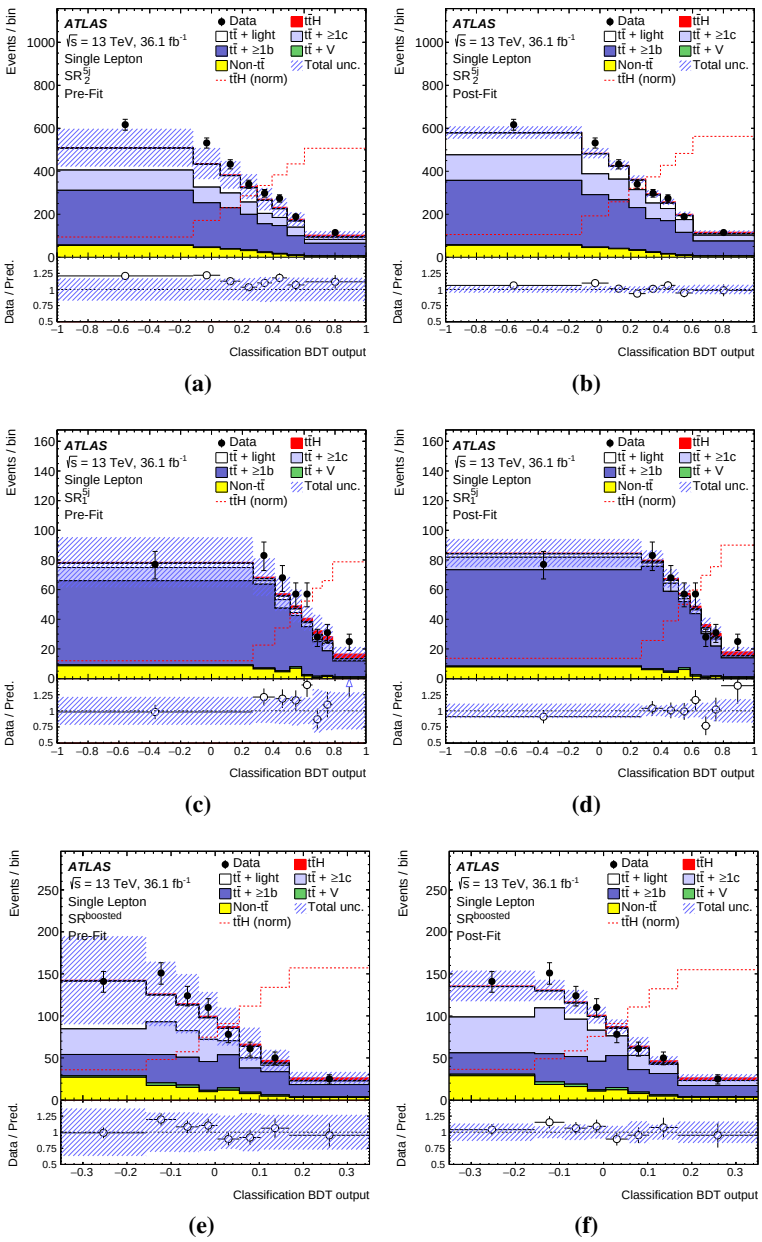


Figure 5.23: Comparison between data and prediction for the classification BDT discriminant in the single-lepton channel five jet and boosted signal regions pre- (left) and post- (right) the combined dilepton and single-lepton fit to data. The $t\bar{t}H$ contribution is normalized to the SM cross section before the fit and to the fitted μ after the fit. The pre-fit plots do not include an uncertainty for the $t\bar{t}+\geq 1b$ or $t\bar{t}+\geq 1c$ normalization.

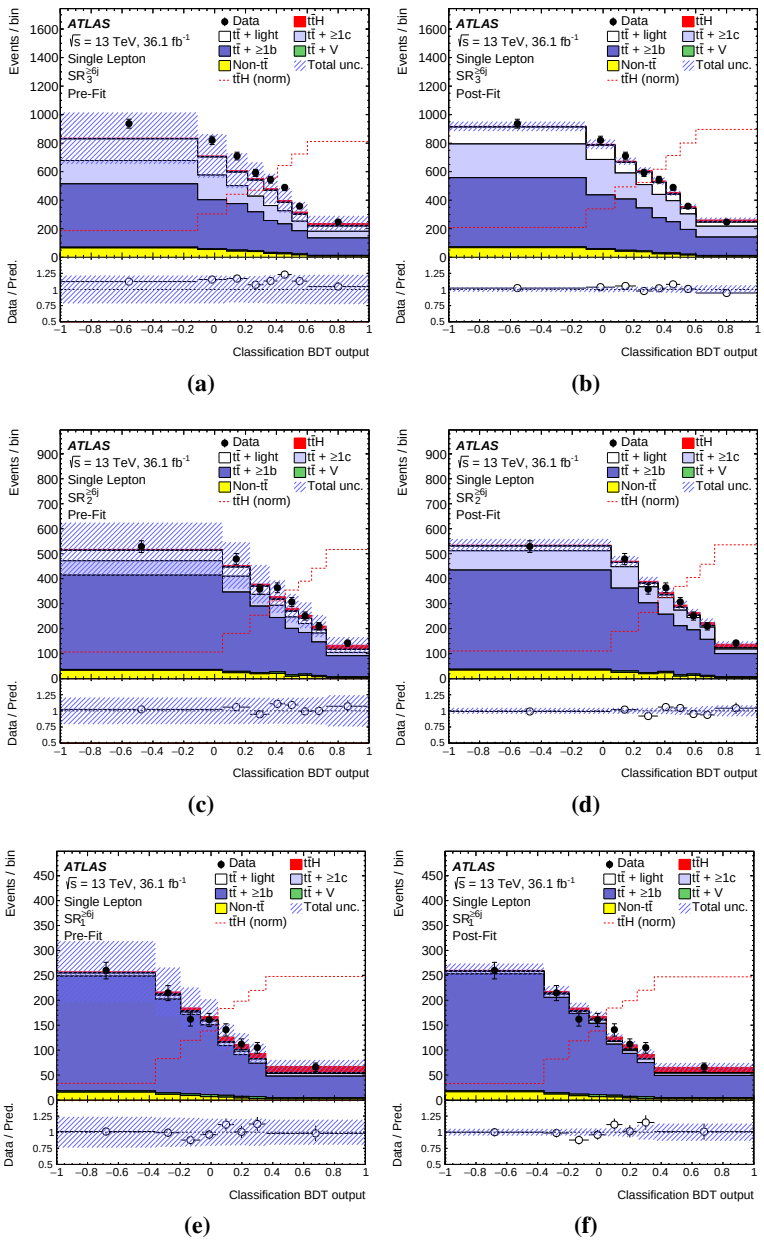


Figure 5.24: Comparison between data and prediction for the BDT discriminant in the single-lepton channel six jet signal regions pre- (left) and post- (right) the combined dilepton and single-lepton fit to data. The $t\bar{t}H$ contribution is normalized to the SM cross section before the fit and to the fitted μ after the fit. The pre-fit plots do not include an uncertainty for the $t\bar{t} + \geq 1b$ or $t\bar{t} + \geq 1c$ normalization.

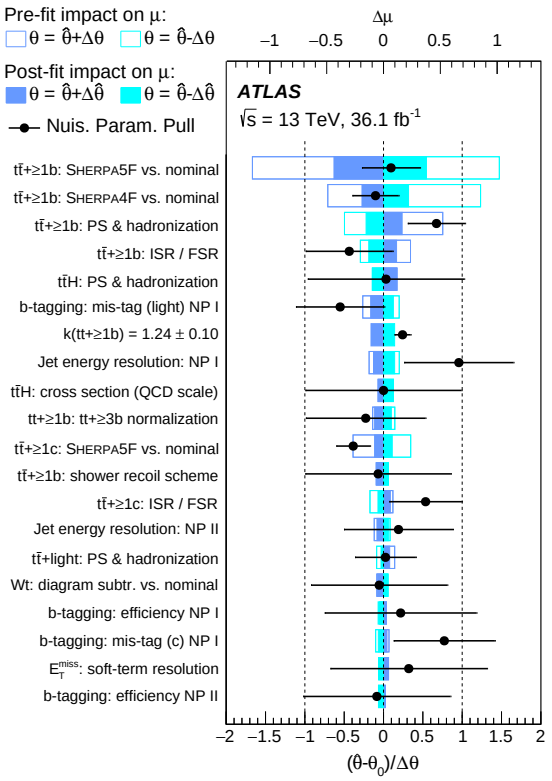


Figure 5.25: Ranking of the top 20 nuisance parameters included in the fit according to their impact on the measured signal strength μ. The empty blue rectangles correspond to the pre-fit impact on μ and the filled blue ones to the post-fit impact on μ, both referring to the upper scale. The black points show the pulls of the nuisance parameters relative to their nominal values, θ_0. These pulls and their relative post-fit errors refer to the scale on the bottom axis. NP I and NP II for experimental uncertainties correspond to the first and second nuisance parameters, ordered by their impact on μ.

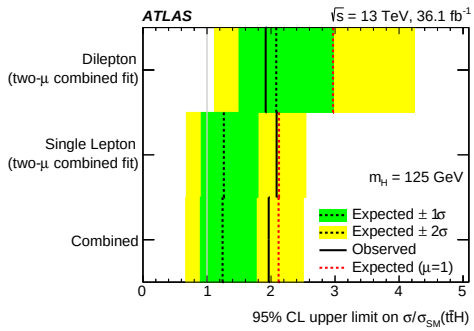


Figure 5.26: Summary of the 95% confidence level (CL) upper limits on $\sigma(t\bar{t}H)$ relative to the SM prediction in the individual channels and for the combination. The observed limits are shown, together with the expected limits both in the background-only hypothesis (dotted black lines) and in the SM hypothesis (dotted red lines). For the background-only hypothesis, one and two standard deviation uncertainty band on the expected limits are shown as well. The limits for the two individual channels are both extracted from the profile likelihood including the data in both channels, but with independent signal strengths in the two channels.

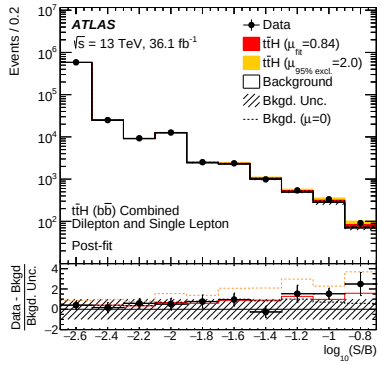


Figure 5.27: Post-fit yields of signal and total background as a function of $\log(S/B)$, compared to data. Final-discriminant bins in all analysis categories are combined into bins of $\log(S/B)$, with the signal normalized to the SM prediction. The red and yellow signal contributions show the signal contribution normalized to the best-fit value and the value excluded at 95% CL. For each bin, the lower panel reports the pull relative to the fitted background prediction. The first bin includes the underflow.

5.7 Combination with other searches

In this section, the combination with the other searches looking for the $t\bar{t}H$ production mode will be briefly presented. In fact, in addition to the results of the $t\bar{t}H(b\bar{b})$ search presented in this thesis, the ATLAS Collaboration has carried out three additional searches for the $t\bar{t}H$ production exploiting other Higgs decay modes:

- $H \to WW^*/\tau\tau/ZZ^* \to$ multi-leptons, with a total of seven final states categorized by the number of charged leptons and their flavours [184], collectively referred to as $H \to$ ML;
- $H \to ZZ^* \to 4\ell$, in a single category including all $t\bar{t}$ decay channels [185];
- $H \to \gamma\gamma$ in lepton+jets, dileptonic and all-hadronic $t\bar{t}$ decay channels [186]. Specialized categories sensitive to $tHqb/WtH$ production also have significant acceptance and are therefore included.

For all the details of the analyses, the reader is referred to the corresponding references. Given that the same 36.1 fb^{-1} dataset collected at $\sqrt{s}$ = 13 TeV was used for all these searches, the overlap among the various analyses has been properly removed and all the cross sections of the non-$t\bar{t}H$ production modes have been fixed to the SM expectations [166].

The combined likelihood is obtained from the product of likelihood functions of the individual analyses. Experimental uncertainties are treated as correlated among the various analyses whenever possible. The cross-section and modelling uncertainties for backgrounds estimated with the MC are correlated between the $H \to b\bar{b}$ and $H \to$ ML analyses, whereas the uncertainties on the dominant $t\bar{t}$ background in $t\bar{t}H(b\bar{b})$ are not correlated with the $t\bar{t}$ modelling of the other analyses, as the relevant phase space and the method used to estimate it are different. Lastly, all analyses use the same nominal Higgs boson production cross sections and decay branching ratios, hence those uncertainties are fully correlated.

The best-fit value of the $t\bar{t}H$ signal strength determined from the com-

bined likelihood function is:

$$\mu = 1.17 \pm 0.19\ (\text{stat.})\ ^{+0.27}_{-0.23}\ (\text{syst.}) \tag{5.17}$$

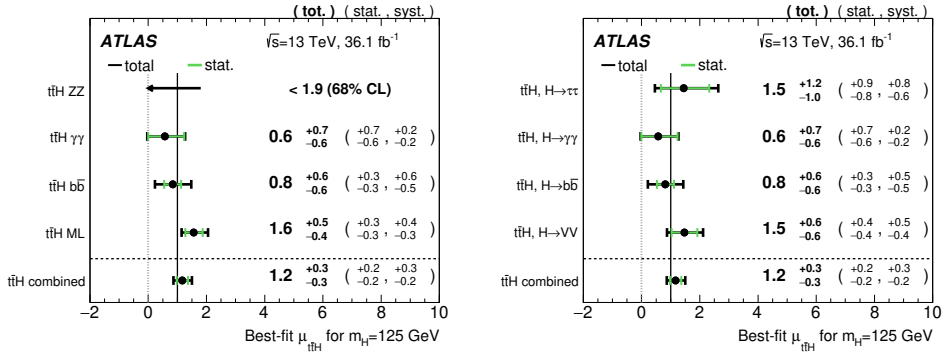
which corresponds to an excess of events over the expected SM background with an observed (expected) significance of 4.2 (3.8) σ, hence the background-only hypothesis is excluded and evidence for the $t\bar{t}H$ production mechanism is found.

The observed signal strengths for the individual analyses and their combination are shown in Figure 5.28a, while Table 5.5 summarizes the observed and expected μ, as well as the significance of $t\bar{t}H$ production from each individual analysis and their combination. The cross section for $t\bar{t}H$ production corresponding to the best-fit value of μ is 590^{+160}_{-150} fb, well compatible with the SM prediction of $\sigma^{\text{SM}}_{t\bar{t}H} = 507^{+35}_{-50}$ fb.

Figure 5.28b shows the result of a fit to four signal strengths, one for the $H \to \tau\tau$, $H \to \gamma\gamma$, $H \to b\bar{b}$ and $H \to VV$ respectively. The category $H \to VV$ combines both $H \to WW^*$ and $H \to ZZ^*$, fixing the ratio of the two branching ratios to its SM prediction, due to very low sensitivity to the latter. Given the high purity of the $H \to b\bar{b}$ and $H \to \gamma\gamma$ for their Higgs boson decay modes, the corresponding signal strengths are substantially the same ones measured in the respective analyses.

Table 5.5: Summary of the observed and expected μ measurements and $t\bar{t}H$ production significance from individual analyses and their combination. As no events are observed in the $H \to 4\ell$ analysis, a 68% confidence level (C.L.) upper limit on μ, computed using the CL_s method is reported.

Channel	Best-fit μ		Significance	
	Observed	Expected	Observed	Expected
$H \to b\bar{b}$	$0.8^{+0.6}_{-0.6}$	$1.0^{+0.4}_{-0.4}$	1.4σ	1.6σ
$H \to \gamma\gamma$	$0.8^{+0.6}_{-0.6}$	$1.0^{+0.8}_{-0.6}$	0.9σ	1.7σ
$H \to 4\ell$	< 1.9	$1.0^{+3.2}_{-1.0}$	–	0.6σ
$H \to$ ML	$1.6^{+0.5}_{-0.4}$	$1.0^{+0.4}_{-0.4}$	4.1σ	2.8σ
Combined	$1.2^{+0.3}_{-0.3}$	$1.0^{+0.3}_{-0.3}$	4.2σ	3.8σ



(a) **(b)**

Figure 5.28: The summary of the measurements of μ from individual analyses and their combination (left) and summary of the best-fit values of μ broken down by Higgs boson decay mode (right) are shown. "ML" refers to the multileptonic decay channels. As no events are observed in the $H \to 4\ell$ analysis, a 68% confidence level (CL) upper limit on μ, computed using the CLs method, is reported. The decays $H \to WW^*$ and $H \to ZZ^*$ are shown together as VV.

The $t\bar{t}H$ analyses are sensitive both to the Higgs to fermion couplings (Htt, Hbb and $H\tau\tau$) and the Higgs to gauge bosons couplings (HWW, HZZ and the effective $H\gamma\gamma$ coupling), therefore constraints can be placed on deviations of these couplings from the SM expectations; the κ-framework [187] is used for making this interpretation.

Within this framework the Higgs coupling to particle i is scaled linearly with a factor κ_i. All fermions are assumed to scale by a common κ_F factor and a common κ_V is employed for the WW/ZZ couplings. Given that only the relative sign of the two κ_i is relevant, κ_V is set ≥ 0 by convention. In loops, only contributions from SM particles are considered. In particular the effective κ_γ factor associated with the $H\gamma\gamma$ coupling is expressed in terms of κ_V and κ_F and κ_g is set equal to κ_F.

There is an interference between the Htt and HWW couplings from the amplitudes of $H \to \gamma\gamma$ decay and the production of the $tHqb$ and WtH. In the SM, this interference is almost completely destructive in

the case of $tHqb$ and WtH, giving the possibility to resolve the relative sign of the two couplings.

A likelihood scan is performed in the κ_F-κ_V plane and the results of this scan, shown in Figure 5.29, are in good agreement with the SM prediction. The possibility that $\kappa_F < 0$ is excluded at 95% CL in this parametrization.

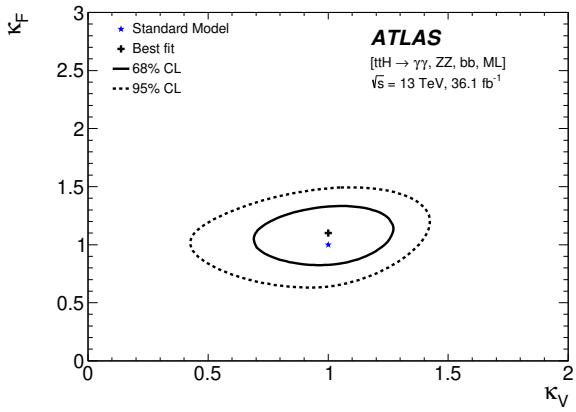


Figure 5.29: Results of the likelihood scan in the κ_F-κ_V plane from the combination of all the $t\bar{t}H$ channels. The solid (dashed) line represents the allowed regions at 68% (95%) CL. The coupling of the Higgs boson with non-SM particles is assumed to be zero, the $H \to \gamma\gamma$ and $H \to gg$ couplings are expressed in terms of κ_V and κ_F.

Conclusions

After the discovery of a new particle compatible with the Higgs boson as predicted by the Standard Model, the focus has shifted on determining its properties, among which there are the Yukawa couplings.

The observation of the decays into fermions and the measurements of the corresponding Yukawa couplings can provide important information about the nature of the newly discovered particle. In this context, the associated production of the Higgs boson with a $t\bar{t}$ pair plays an important role, as the $t\bar{t}H$ channel is the only one that allows a direct measurement of this coupling at the LHC.

The core of this thesis is the search for the Standard Model Higgs boson produced in association with a pair of top quarks and decaying into a $b\bar{b}$ pair. The results presented in Chapter 5 are obtained analysing the data collected by the ATLAS experiment during 2015 and 2016, corresponding to 36.1 fb^{-1} of integrated luminosity.

The topology of the $t\bar{t}$ decay is used to define the analysis channels: resolved and boosted single-lepton, and dilepton channel. In order to reduce the importance of the overwhelming $t\bar{t} + \geq 1b$ background, a complex strategy employing multivariate techniques is put in place.

Selected events are first categorized into signal- and background-enriched regions. Two layers of multivariate techniques then are used in the signal regions: the first one is used to reconstruct the final state, while the second is used to classify events into signal- and background-like.

Finally, a profile likelihood fit to data is used to measure the signal strength, whose best-fit value in all the single-lepton and dilepton regions yields a value of:

$$\mu = 0.84 \pm 0.29\ (\text{stat.})\ ^{+0.57}_{-0.54}\ (\text{syst.}) = 0.84^{+0.64}_{-0.61}$$

with the expected uncertainty identical to the measured one. An excess of events over the expected SM background is found with an observed (expected) significance of 1.4 (1.6) standard deviations.

Recent updates on the combinations with other searches

Other analyses searching for $t\bar{t}H$ production and exploiting different decay modes have been performed in ATLAS: the Higgs boson decaying into a pair of photons, $t\bar{t}H(\to \gamma\gamma)$, the Higgs boson decay into four leptons, $t\bar{t}H(\to ZZ^* \to 4\ell)$, and the decay into a multi-leptons final state, $t\bar{t}H(\to WW^*/\tau\tau/ZZ^* \to \text{leptons})$. The combination of these analyses, performed on 36.1 fb^{-1} of data, was presented in Section 5.7.

This section presents the most recent updates on that combination. Compared to those results, the $t\bar{t}H(\to \gamma\gamma)$ and $t\bar{t}H(\to ZZ^* \to 4\ell)$ analyses are updated with the 13 TeV data collected during 2017 for a total of 79.8 fb^{-1}. In addition, improved photon and lepton reconstruction algorithms and analysis techniques are used.

The $t\bar{t}H(b\bar{b})$ and $t\bar{t}H(\to WW^*/\tau\tau/ZZ^* \to \text{leptons})$ analyses done with 36.1 fb^{-1} were combined with these updated analyses; the combined likelihood fit to extract the $t\bar{t}H$ signal results in an observed (expected) excess relative to the background-only hypothesis of 5.8 (4.9) standard deviations, therefore the Higgs boson production in association with a top quark pair has been observed with the ATLAS detector [188].

The measured total production cross section for the $t\bar{t}H$ process at 13 TeV is $670 \pm 90(\text{stat.})^{+110}_{-100}(\text{syst.})$ fb, in reasonable agreement with the SM prediction of $\sigma^{\text{SM}}_{t\bar{t}H} = 507^{+35}_{-50}$ fb. Figure 1 shows the ratios of the cross sections extracted in the combined likelihood fit to their SM predictions.

In ATLAS, other searches targeting the $H \to b\bar{b}$ decay mode have been performed; among them, the VH($\to bb$) analysis stands out for its high significance.

The search for the SM VH($\to bb$) process was performed using data collected at $\sqrt{s} = 13$ TeV for a total of 79.8 fb^{-1} of integrated luminosity and the results were combined with the searches for the $t\bar{t}H(b\bar{b})$ and vector-boson fusion for both the Run1 and Run2.

Assuming a Higgs boson mass of 125 GeV and assuming that the

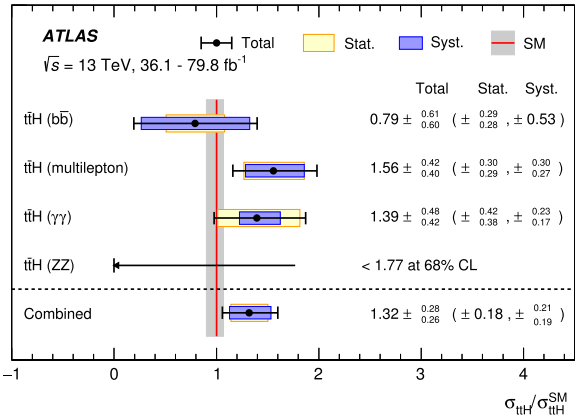


Figure 1: Combined $t\bar{t}H$ production cross section, as well as cross sections measured in the individual analyses, divided by the SM prediction. The $\gamma\gamma$ and $ZZ^* \to 4\ell$ analyses use 13 TeV data corresponding to an integrated luminosity of 79.8 fb^{-1}, while the multilepton and $b\bar{b}$ analyses use data corresponding to an integrated luminosity of 36.1 fb^{-1}. The black lines show the total uncertainties and the bands indicate the statistical and systematic uncertainties. The red vertical line indicates the SM cross-section prediction and the grey band represents its associated uncertainty.

production cross-sections are those predicted by the SM, the observed significance for the $H \to b\bar{b}$ decay is 5.4 σ, to be compared with the expected value of 5.5 standard deviations. Figure 2 shows the signal strengths obtained from a fit where individual signal strengths are simultaneously fitted for the three production modes displayed.

Outlook

All the recent results look consistent with the Standard Model predictions. Nevertheless, the uncertainties are large and improved measurements will come in the future due to both an increase in the amount of data collected by the end of Run2 and in the reduction of the systematic uncertainties.

The uncertainties on the results of the $t\bar{t}H(b\bar{b})$ analysis are already

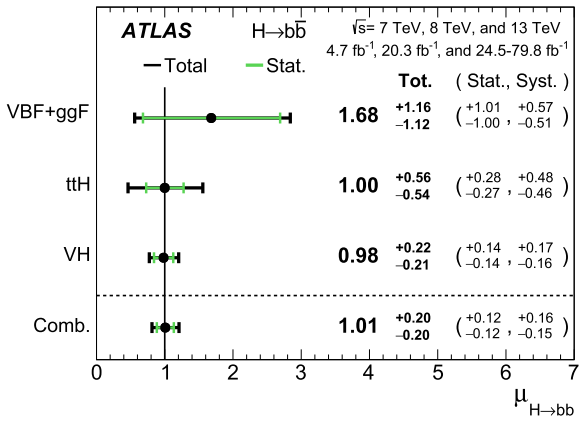


Figure 2: The fitted values of the Higgs boson signal strength $\mu_{H\to bb}$ for $m_H = 125$ GeV separately for the VH, $t\bar{t}H$ and and VBF+ggF analyses along with their combination, using the 7 TeV, 8 TeV and 13 TeV data. The individual $\mu_{H\to bb}$ values for the different production modes are obtained from a simultaneous fit with the signal strengths for each of the processes floating independently.

dominated by the systematic component. Improvements of the $t\bar{t}$ + jets background modelling, and its $t\bar{t} + \geq 1b$ component, are of vital importance in order to improve the precision of the analysis. In this respect, dedicated measurements of the three main components are desirable in order to reduce differences among the various Monte Carlo generators available and reduce the uncertainties on the $t\bar{t}$ + jets prediction.

A second important ingredient, which is fairly easy to put in place, is the increase in the amount of MC events generated in the phase space selected by the analysis, as this will reduce the statistical fluctuations affecting the background predictions, which are the second largest source of error on the measured signal strength.

All these desirable improvements are not likely to occur in a short timescale, for this reason the most precise determination of the top Yukawa coupling is expected to be driven by the $t\bar{t}H(\to \gamma\gamma)$ and the $t\bar{t}H(\to$ leptons$)$ searches, which will have a sensitivity higher than the $t\bar{t}H(b\bar{b})$ one, due to a smaller impact of their systematic uncertainties

and the extremely high purity.

Projections for measurements of Higgs boson signal strengths and coupling parameters have been performed, using 14 TeV proton-proton collisions at the LHC with 300 fb^{-1} and at the High-Luminosity LHC (HL-LHC) with 3000 fb^{-1} [189]. This prospect study combines several Higgs decay channels, among which the $t\bar{t}H(\rightarrow \gamma\gamma)$ but not the $t\bar{t}H(b\bar{b})$ one, and shows that the relative uncertainty on the signal strength should be about 30% with 300 fb^{-1} and 10-15% with 3000 fb^{-1}, allowing for a precise measurements of the top Yukawa coupling. Furthermore, the cross-section of the dominant ggF production mode should reach an experimental precision of about 4%, close to the limit given by the assumed luminosity uncertainty of 3%, providing strong constraint on possible BSM contributions to the loop diagrams.

Bibliography

[1] S. L. Glashow. *Partial Symmetries of Weak Interactions*. Nucl. Phys. **22** (1961), pp. 579–588.

[2] S. Weinberg. *A Model of Leptons*. Phys. Rev. Lett. **19** (1967), pp. 1264–1266.

[3] A. Salam. *Weak and Electromagnetic Interactions*. 8th Nobel Symposium Lerum, Sweden (1968), pp. 367–377.

[4] R. P. Feynman. *Very High-Energy Collisions of Hadrons*. Phys. Rev. Lett. **23** (1969), pp. 1415–1417.

[5] M. Gell-Mann. *A Schematic Model of Baryons and Mesons*. Phys. Lett. **8** (1964), pp. 214–215.

[6] G. Zweig. *An SU_3 model for strong interaction symmetry and its breaking; Version 1* (1964).

[7] F. Mandl and G. Shaw. *Quantum Field Theory*. Ed. by Wiley. 2010. ISBN: 978-0-471-49683-0.

[8] Particle Data Group. *Review of Particle Physics*. Phys. Rev. D **98** (2018), p. 030001.

[9] SNO Collaboration. *Direct evidence for neutrino flavor transformation from neutral current interactions in the Sudbury Neutrino Observatory*. Phys. Rev. Lett. **89** (2002), p. 011301. arXiv: `nucl-ex/0204008 [nucl-ex]`.

[10] Super-Kamiokande Collaboration. *Evidence for an oscillatory signature in atmospheric neutrino oscillation*. Phys. Rev. Lett. **93** (2004), p. 101801. arXiv: `hep-ex/0404034 [hep-ex]`.

[11] E. Noether. *Invariante Variationsprobleme (Invariant variational problems)*. Gott.Nachr. (1918), pp. 235–257.

[12] D. Hanneke, S. Fogwell and G. Gabrielse. *New Measurement of the Electron Magnetic Moment and the Fine Structure Constant*. Phys. Rev. Lett. **100** (2008), p. 120801.

[13] R. Bouchendira, P. Cladé, S. Guellati-Khélifa, F. Nez and F. Biraben. *New Determination of the Fine Structure Constant and Test of the Quantum Electrodynamics*. Phys. Rev. Lett. **106** (2011), p. 080801.

[14] C. N. Yang and R. L. Mills. *Conservation of Isotopic Spin and Isotopic Gauge Invariance*. Phys. Rev. **96** (1954), pp. 191–195.

[15] Y. Nambu. *Quasi-Particles and Gauge Invariance in the Theory of Superconductivity*. Phys. Rev. **117** (1960), pp. 648–663.

[16] J. Goldstone. *Field theories with « Superconductor » solutions*. Nuovo Cimento **19** (1961), pp. 154–164.

[17] J. Goldstone, A. Salam and S. Weinberg. *Broken Symmetries*. Phys. Rev. **127** (1962), pp. 965–970.

[18] F. Englert and R. Brout. *Broken Symmetry and the Mass of Gauge Vector Mesons*. Phys. Rev. Lett. **13** (1964), pp. 321–323.

[19] P. Higgs. *Broken Symmetries and the Masses of Gauge Bosons*. Phys. Rev. Lett. **13** (1964), pp. 508–509.

[20] P. Higgs. *Broken symmetries, massless particles and gauge fields*. Phys. Lett. **12** (1964), pp. 132–133.

[21] G. S. Guralnik, C. R. Hagen and T. W. B. Kibble. *Global Conservation Laws and Massless Particles*. Phys. Rev. Lett. **13** (1964), pp. 585–587.

[22] G. S. Guralnik. *The History of the Guralnik, Hagen and Kibble development of the Theory of Spontaneous Symmetry Breaking and Gauge Particles*. Int. J. Mod. Phys. A **24** (2009), pp. 2601–2627. arXiv: 0907.3466 [physics.hist-ph].

[23] P. Sikivie, L. Susskind, M. Voloshin and V. Zakharov. *Isospin breaking in technicolor models*. Nucl. Phys. B **173** (1980), pp. 189–207.

[24] N. Cabibbo. *Unitary Symmetry and Leptonic Decays*. Phys. Rev. Lett. **10** (1963), pp. 531–533.

[25] M. Kobayashi and T. Maskawa. *CP-Violation in the Renormalizable Theory of Weak Interaction*. **49** (1973), pp. 652–657.

[26] L. Wolfenstein. *Parametrization of the Kobayashi–Maskawa Matrix*. Phys. Rev. Lett. **51** (1983), pp. 1945–1947.

[27] V. N. Gribov and L. N. Lipatov. *Deep inelastic ep scattering in perturbation theory*. Sov. J. Nucl. Phys. **15** (1972), pp. 438–450.

[28] L. N. Lipatov. *The parton model and perturbation theory*. Sov. J. Nucl. Phys. **20** (1975), pp. 94–102.

[29] G. Altarelli and G. Parisi. *Asymptotic Freedom in Parton Language*. Nucl. Phys. B **126** (1977), pp. 298–318.

[30] Y. L. Dokshitzer. *Calculation of the Structure Functions for Deep Inelastic Scattering and e+ e- Annihilation by Perturbation Theory in Quantum Chromodynamics.* Sov. Phys. JETP **46** (1977), pp. 641–653.

[31] LHC Higgs Cross Section Working Group. *Handbook of LHC Higgs Cross Sections: 1. Inclusive Observables* (2011). arXiv: 1101.0593 [hep-ph].

[32] LHC Higgs Cross Section Working Group. *Handbook of LHC Higgs Cross Sections: 3. Higgs Properties* (2013). Ed. by S. Heinemeyer, C. Mariotti, G. Passarino and R. Tanaka. arXiv: 1307.1347 [hep-ph].

[33] UA1 Collaboration. *Experimental observation of isolated large transverse energy electrons with associated missing energy at s=540 GeV*. Phys. Lett. B **122** (1983), pp. 103–116.

[34] UA2 Collaboration. *Observation of single isolated electrons of high transverse momentum in events with missing transverse energy at the CERN pp collider*. Phys. Lett. B **122** (1983), pp. 476–485.

[35] D0 Collaboration. *Search for High Mass Top Quark Production in $p\overline{p}$ Collisions at $\sqrt{s} = 1.8$ TeV*. Phys. Rev. Lett. **74** (1995), pp. 2422–2426.

[36] CDF Collaboration. *Observation of Top Quark Production in $p\overline{p}$ Collisions with the Collider Detector at Fermilab*. Phys. Rev. Lett. **74** (1995), pp. 2626–2631.

[37] M. Baak et al. *Updated Status of the Global Electroweak Fit and Constraints on New Physics*. Eur. Phys. J. C **72** (2012), p. 2003. arXiv: 1107.0975 [hep-ph].

[38] ALEPH, DELPHI, L3, OPAL Collaborations, the LEP Working Group for Higgs boson searches. *Search for the standard model Higgs boson at LEP*. Phys. Lett. B **565** (2003), pp. 61–75. arXiv: hep-ex/0306033 [hep-ex].

[39] The TEVNPH Working Group, for the CDF, D0 Collaborations. *Combined CDF and D0 Search for Standard Model Higgs Boson Production with up to 10.0 fb^{-1} of Data* (2012). arXiv: 1203.3774 [hep-ex].

[40] ATLAS Collaboration. *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*. Phys. Lett. B **716** (2012), pp. 1–29. arXiv: 1207.7214 [hep-ex].

[41] CMS Collaboration. *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*. Phys. Lett. B **716** (2012), pp. 30–61.

[42] ATLAS, CMS Collaborations. *Combined Measurement of the Higgs Boson Mass in pp Collisions at $\sqrt{s} = 7$ and 8 TeV with the ATLAS and CMS Experiments*. Phys. Rev. Lett. **114** (2015), p. 191803. arXiv: 1503.07589 [hep-ex].

[43] ATLAS and CMS Collaborations. *Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysis of the LHC pp collision data at $\sqrt{s} = 7$ and 8 TeV*. JHEP **08** (2016), p. 045. arXiv: 1606.02266 [hep-ex].

[44] ATLAS Collaboration. *Search for the Standard Model Higgs boson decaying into $b\bar{b}$ produced in association with top quarks decaying hadronically in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector*. JHEP **05** (2016), p. 160. arXiv: 1604.03812 [hep-ex].

[45] ATLAS Collaboration. *Summary plots from the ATLAS Standard Model physics group*. URL: https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/CombinedSummaryPlots/SM/.

[46] A. Riotto. *Theories of baryogenesis*. In: *Proceedings, Summer School in High-energy physics and cosmology: Trieste, Italy, June 29-July 17, 1998*, pp. 326–436. arXiv: hep-ph/9807454 [hep-ph].

[47] A. V. Bednyakov, B. A. Kniehl, A. F. Pikelner and O. L. Veretin. *Stability of the Electroweak Vacuum: Gauge Independence and Advanced Precision*. Phys. Rev. Lett. **115** (2015), p. 201802.

[48] F. Bezrukov and M. Shaposhnikov. *Why should we care about the top quark Yukawa coupling?* J. Exp. Theor. Phys. **120** (2015), pp. 335–343. arXiv: 1411.1923 [hep-ph].

[49] L. Evans and P. Bryant. *LHC Machine*. JINST **3** (2008), S08001.

[50] ATLAS Collaboration. *ATLAS: Technical proposal for a general purpose pp experiment at the Large Hadron Collider at CERN* (1994).

[51] ATLAS Collaboration. *The ATLAS Experiment at the CERN Large Hadron Collider*. JINST **3** (2008), S08003. 437 p.

[52] CMS Collaboration. *The CMS Experiment at the CERN LHC*. JINST **3** (2008), S08004.

[53] LHCb Collaboration. *LHCb : Technical Proposal* (1998).

[54] ALICE Collaboration. *ALICE: Technical proposal for a Large Ion collider Experiment at the CERN LHC* (1995).

[55] TOTEM Collaboration. *TOTEM: Technical design report. Total cross section, elastic scattering and diffraction dissociation at the Large Hadron Collider at CERN* (2004).

[56] LHCf Collaboration. *Technical design report of the LHCf experiment: Measurement of photons and neutral pions in the very forward region of LHC* (2006).

[57] F. Marcastel. *CERN's Accelerator Complex. La chaîne des accélérateurs du CERN*. 2013. URL: https://cds.cern.ch/record/1621583.

[58] ATLAS Collaboration. *ATLAS luminosity public results in Run2*. URL: https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResultsRun2.

[59] ATLAS Collaboration. *ATLAS magnet system: Technical design report* (1997).

[60] J. Pequenao. *Computer generated image of the whole ATLAS detector*. 2008. URL: https://cds.cern.ch/record/1095924.

[61] J. Pequenao. *Computer generated image of the ATLAS inner detector*. 2008. URL: https://cds.cern.ch/record/1095926.

[62] ATLAS Collaboration. *ATLAS inner detector: Technical design report. Vol. 1* (1997).

[63] ATLAS Collaboration. *Track Reconstruction Performance of the ATLAS Inner Detector at* $\sqrt{s} = 13$ *TeV*. ATL-PHYS-PUB-2015-018. URL: https://cds.cern.ch/record/2037683.

[64] ATLAS Collaboration. *ATLAS Insertable B-Layer Technical Design Report*. ATLAS-TDR-19. 2010. URL: https://cds.cern.ch/record/1291633.

[65] ATLAS Collaboration. *IBL Efficiency and Single Point Resolution in Collision Events*. ATL-INDET-PUB-2016-001. URL: https://cds.cern.ch/record/2203893.

[66] J. Pequenao. *Computer Generated image of the ATLAS calorimeter*. 2008. URL: https://cds.cern.ch/record/1095927.

[67] ATLAS Collaboration. *ATLAS liquid argon calorimeter: Technical design report* (1996).

[68] ATLAS Collaboration. *ATLAS tile calorimeter: Technical design report* (1996).

[69] J. Pequenao. *Computer generated image of the ATLAS Muons subsystem*. 2008. URL: http://cds.cern.ch/record/1095929.

[70] ATLAS Collaboration. *ATLAS muon spectrometer: Technical design report* (1997).

[71] ATLAS Collaboration. *Performance of the ATLAS muon trigger in pp collisions at $\sqrt{s} = 8$ TeV*. Eur. Phys. J. C **75** (2015), p. 120. arXiv: 1408.3179 [hep-ex].

[72] ATLAS Collaboration. *Technical Design Report for the Phase-I Upgrade of the ATLAS TDAQ System*. CERN-LHCC-2013-018, ATLAS-TDR-023. Final version presented to December 2013 LHCC.

[73] ATLAS Collaboration. *Performance of the ATLAS trigger system in 2015*. Eur. Phys. J. C **77** (2017), p. 317.

[74] S. van der Meer. *Calibration of the effective beam height in the ISR*. 1968. URL: https://cds.cern.ch/record/296752.

[75] ATLAS Collaboration. *Luminosity determination in pp collisions at $\sqrt{s}$ = 8 TeV using the ATLAS detector at the LHC*. Eur. Phys. J. C **76** (2016), p. 653. arXiv: 1608.03953 [hep-ex].

[76] ATLAS Collaboration. *The ATLAS Simulation Infrastructure*. Eur. Phys. J. C **70** (2010), p. 823. arXiv: 1005.4568 [physics.ins-det].

[77] S. Agostinelli et al. *GEANT4: A Simulation toolkit*. Nucl. Instrum. Meth. A **506** (2003), pp. 250–303.

[78] W. Lukas. *Fast Simulation for ATLAS: Atlfast-II and ISF*. 2012. URL: https://cds.cern.ch/record/1458503.

[79] T. Gleisberg et al. *Event generation with SHERPA 1.1*. JHEP **02** (2009), p. 007. arXiv: 0811.4622 [hep-ph].

[80] J. Bellm et al. *Herwig 7.0/Herwig++ 3.0 release note*. Eur. Phys. J. C **76** (2016), p. 196. arXiv: 1512.01178 [hep-ph].

[81] M. Bähr et al. *Herwig++ physics and manual*. Eur. Phys. J. C **58** (2008), pp. 639–707. arXiv: 0803.0883 [hep-ph].

[82] T. Sjostrand et al. *An Introduction to PYTHIA 8.2*. Comput. Phys. Commun. **191** (2015), pp. 159–177. arXiv: 1410.3012 [hep-ph].

[83] J. Alwall et al. *The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations*. JHEP **07** (2014), p. 079. arXiv: 1405.0301 [hep-ph].

[84] P. Nason. *A New method for combining NLO QCD with shower Monte Carlo algorithms*. JHEP **11** (2004), p. 040. arXiv: hep-ph/0409146.

[85] S. Frixione, P. Nason and C. Oleari. *Matching NLO QCD computations with Parton Shower simulations: the POWHEG method*. JHEP **11** (2007), p. 070. arXiv: 0709.2092 [hep-ph].

[86] S. Alioli, P. Nason, C. Oleari and E. Re. *A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX*. JHEP **06** (2010), p. 043. arXiv: 1002.2581 [hep-ph].

[87] J. M. Campbell, R. K. Ellis, P. Nason and E. Re. *Top-pair production and decay at NLO matched with parton showers*. JHEP **04** (2015), p. 114. arXiv: 1412.1828 [hep-ph].

[88] ATLAS Collaboration. *ATLAS Run 1 Pythia8 tunes*. ATL-PHYS-PUB-2014-021. URL: https://cds.cern.ch/record/1966419.

[89] P. Z. Skands. *Tuning Monte Carlo generators: The Perugia tunes*. Phys. Rev. D **82** (2010), p. 074018. arXiv: 1005.3457 [hep-ph].

[90] T. Cornelissen et al. *The new ATLAS track reconstruction (NEWT)*. J. Phys. Conf. Ser. **119** (2008), p. 032014.

[91] A. Salzburger. *Optimisation of the ATLAS Track Reconstruction Software for Run-2*. J. Phys. Conf. Ser. **664** (2015), p. 072042.

[92] R. Fruhwirth. *Application of Kalman filtering to track and vertex fitting*. Nucl. Instrum. Meth. A **262** (1987), pp. 444–450.

[93] J. Pequenao. *Event Cross Section in a computer generated image of the ATLAS detector.* 2008. URL: https://cds.cern.ch/record/1096081.

[94] ATLAS Collaboration. *Early Inner Detector Tracking Performance in the 2015 data at $\sqrt{s}$ = 13 TeV*. ATL-PHYS-PUB-2015-051. URL: https://cds.cern.ch/record/2110140.

[95] ATLAS Collaboration. *Vertex Reconstruction Performance of the ATLAS Detector at $\sqrt{s}$ = 13 TeV*. ATL-PHYS-PUB-2015-026. URL: https://cds.cern.ch/record/2037717.

[96] ATLAS Collaboration. *Measurement of the tau lepton reconstruction and identification performance in the ATLAS experiment using pp collisions at $\sqrt{s} = 13$ TeV*. ATLAS-CONF-2017-029. URL: https://cds.cern.ch/record/2261772.

[97] ATLAS Collaboration. *Electron efficiency measurements with the ATLAS detector using the 2015 LHC proton-proton collision data*. ATLAS-CONF-2016-024. URL: https://cds.cern.ch/record/2157687.

[98] ATLAS Collaboration. *Electron and photon energy calibration with the ATLAS detector using data collected in 2015 at $\sqrt{s} = 13$ TeV*. ATL-PHYS-PUB-2016-015. URL: https://cds.cern.ch/record/2203514.

[99] ATLAS Collaboration. *Muon reconstruction performance of the ATLAS detector in proton–proton collision data at $\sqrt{s} = 13$ TeV*. Eur. Phys. J. C **76** (2016), p. 292. arXiv: 1603.05598 [hep-ex].

[100] W. Lampl et al. *Calorimeter clustering algorithms: Description and performance* (2008).

[101] M. Cacciari, G. P. Salam and G. Soyez. *The anti-k_t jet clustering algorithm*. JHEP (2008), p. 063. arXiv: 0802.1189 [hep-ph].

[102] G. P. Salam. *Towards Jetography*. Eur. Phys. J. C **67** (2010), pp. 637–686. arXiv: 0906.1833 [hep-ph].

[103] ATLAS Collaboration. *Jet energy scale measurements and their systematic uncertainties in proton-proton collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector*. Phys. Rev. D **96** (2017), p. 072002.

[104] ATLAS Collaboration. *Jet energy measurement and its systematic uncertainty in proton–proton collisions at $\sqrt{s}$ = 7 TeV with the ATLAS detector*. Eur. Phys. J. C **75** (2015), p. 17.

[105] M. Cacciari and G. P. Salam. *Pileup subtraction using jet areas*. Phys. Lett. B **659** (2008), pp. 119–126.

[106] ATLAS Collaboration. *Jet global sequential corrections with the ATLAS detector in proton-proton collisions at $\sqrt{s}$ = 8 TeV*. ATLAS-CONF-2015-002. URL: https://cds.cern.ch/record/2001682.

[107] ATLAS Collaboration. *Properties of Jets and Inputs to Jet Reconstruction and Calibration with the ATLAS Detector Using Proton-Proton Collisions at $\sqrt{s} = 13$ TeV*. ATL-PHYS-PUB-2015-036. URL: https://cds.cern.ch/record/2044564.

[108] ATLAS Collaboration. *Selection of jets produced in 13 TeV proton-proton collisions with the ATLAS detector*. ATLAS-CONF-2015-029. URL: https://cds.cern.ch/record/2037702.

[109] ATLAS Collaboration. *Performance of pile-up mitigation techniques for jets in pp collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector*. Eur. Phys. J. C **76** (2016), p. 581.

[110] ATLAS Collaboration. *Secondary vertex finding for jet flavour identification with the ATLAS detector*. ATL-PHYS-PUB-2017-011. URL: https://cds.cern.ch/record/2270366.

[111] ATLAS Collaboration. *Plots of b-tagging performance before and after the installation of the Insertable B-Layer*. URL: http://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/FTAG-2017-005/.

[112] G. Piacquadio and C. Weiser. *A new inclusive secondary vertex algorithm for b-jet tagging in ATLAS*. J. Phys. Conf. Ser. **119** (2008), p. 032032.

[113] ATLAS Collaboration. *Expected performance of the ATLAS b-tagging algorithms in Run-2*. ATL-PHYS-PUB-2015-022. URL: https://cds.cern.ch/record/2037697.

[114] ATLAS Collaboration. *Optimisation of the ATLAS b-tagging performance for the 2016 LHC Run*. ATL-PHYS-PUB-2016-012. URL: https://cds.cern.ch/record/2160731.

[115] ATLAS Collaboration. *Measurements of b-jet tagging efficiency with the ATLAS detector using $t\bar{t}$ events at $\sqrt{s}$ = 13 TeV*. JHEP **08** (2018), p. 89. arXiv: 1805.01845 [hep-ex].

[116] ATLAS Collaboration. *E_T^{miss} performance in the ATLAS detector using 2015-2016 LHC pp collisions*. ATLAS-CONF-2018-023. URL: https://cds.cern.ch/record/2625233.

[117] R.D. Field and R.P. Feynman. *A parametrization of the properties of quark jets*. Nucl. Phys. B **136** (1978), pp. 1–76.

[118] ATLAS Collaboration. *Measurement of jet charge in dijet events from $\sqrt{s}$ =8 TeV pp collisions with the ATLAS detector*. ATLAS-CONF-2015-025. URL: https://cdsweb.cern.ch/record/2037618.

[119] ATLAS Collaboration. *Jet Charge Studies with the ATLAS Detector Using $\sqrt{s} = 8$ TeV Proton–Proton Collision Data*. ATLAS-CONF-2013-086. URL: https://cds.cern.ch/record/1572980.

[120] ATLAS Collaboration. *Flavor tagged time-dependent angular analysis of the $B_s \to J/\psi\phi$ decay and extraction of $\Delta\Gamma_s$ and the weak phase ϕ_s in ATLAS*. Phys. Rev. D **90** (2014), p. 052007. arXiv: 1407.1796 [hep-ex].

[121] T. Jakoubek. *Flavour tagging techniques for CPV studies in the B_s system with ATLAS* (2014). arXiv: 1410.8732 [hep-ex].

[122] ATLAS Collaboration. *Measurement of the top quark charge in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector*. JHEP **1311** (2013), p. 031. arXiv: 1307.4568 [hep-ex].

[123] K. Abe et al. *Direct Measurements of A_b and A_c Using Vertex and Kaon Charge Tags at the SLAC Detector*. Phys. Rev. Lett. **94** (2005), p. 091801.

[124] ATLAS Collaboration. *A new tagger for the charge identification of b-jets*. ATL-PHYS-PUB-2015-040. URL: https://cds.cern.ch/record/2048132.

[125] ATLAS Collaboration. *Measurement of the Jet Vertex Charge algorithm performance for identified b-jets in $t\bar{t}$ events in pp collisions with the ATLAS detector*. ATLAS-CONF-2018-022. URL: https://cds.cern.ch/record/2622370.

[126] A. L. Samuel. *Some Studies in Machine Learning Using the Game of Checkers*. IBM Journal of Research and Development **3** (1959), pp. 210–229.

[127] DeepMind. *AlphaGo Home*. 2017. URL: `https://deepmind.com/research/alphago/`.

[128] Silver David et al. *Mastering the game of Go with deep neural networks and tree search*. Nature **529** (2016), pp. 484–489.

[129] O. Behnke, K. Kröninger, T. Schörner-Sadenius and G. Schott. *Data analysis in high energy physics: a practical guide to statistical methods*. Ed. by Wiley. 2013. ISBN: 978-3-527-41058-3.

[130] A. Hoecker et al. *TMVA: Toolkit for Multivariate Data Analysis*. PoS **ACAT** (2007), p. 040. arXiv: `physics/0703039`.

[131] J. Neyman and E. Pearson. *On the problem of the most efficient tests of statistical hypotheses*. Philosophical Transactions of the Royal Society of London A **231** (1933), pp. 289–337.

[132] D. J. Lange. *The EvtGen particle decay simulation package*. Nucl. Instrum. Meth. A **462** (2001), pp. 152–155.

[133] M. Czakon and A. Mitov. *Top++: a program for the calculation of the top-pair cross-section at hadron colliders*. Comput. Phys. Commun. **185** (2014), p. 2930. arXiv: `1112.5675 [hep-ph]`.

[134] M. Cacciari, M. Czakon, M. Mangano, A. Mitov and P. Nason. *Top-pair production at hadron colliders with next-to-next-to-leading logarithmic soft-gluon resummation*. Phys. Lett. B **710** (2012), p. 612. arXiv: `1111.5869 [hep-ph]`.

[135] M. Czakon and A. Mitov. *NNLO corrections to top-pair production at hadron colliders: the all-fermionic scattering channels*. JHEP **12** (2012), p. 054. arXiv: `1207.0236 [hep-ph]`.

[136] M. Czakon and A. Mitov. *NNLO corrections to top pair production at hadron colliders: the quark-gluon reaction*. JHEP **01** (2013), p. 080. arXiv: `1210.6832 [hep-ph]`.

[137] M. Czakon, P. Fiedler and A. Mitov. *Total Top-Quark Pair-Production Cross Section at Hadron Colliders Through $O(\alpha_S^4)$*. Phys. Rev. Lett. **110** (2013), p. 252004. arXiv: `1303.6254 [hep-ph]`.

[138] P. Barnreuther, M. Czakon and A. Mitov. *Percent-Level-Precision Physics at the Tevatron: Next-to-Next-to-Leading Order QCD Corrections to $q\bar{q} \to t\bar{t}+X$*. Phys. Rev. Lett. **109** (2012), p. 132001.

[139] ATLAS Collaboration. *Studies on top-quark Monte Carlo modelling for Top2016*. ATL-PHYS-PUB-2016-020. URL: https://cds.cern.ch/record/2216168.

[140] ATLAS Collaboration. *Simulation of top-quark production for the ATLAS experiment at $\sqrt{s} = 13$ TeV*. ATL-PHYS-PUB-2016-004. URL: https://cds.cern.ch/record/2120417.

[141] N. Kidonakis. *Two-loop soft anomalous dimensions for single top quark associated production with a W- or H-*. Phys. Rev. D **82** (2010), p. 054018. arXiv: 1005.4451 [hep-ph].

[142] N. Kidonakis. *NNLL resummation for s-channel single top quark production*. Phys. Rev. D **81** (2010), p. 054028. arXiv: 1001.5034 [hep-ph].

[143] N. Kidonakis. *Next-to-next-to-leading-order collinear and soft gluon corrections for t-channel single top quark production*. Phys. Rev. D **83** (2011), p. 091503. arXiv: 1103.2792 [hep-ph].

[144] T. Sjostrand, S. Mrenna and P. Z. Skands. *PYTHIA 6.4 physics and manual*. JHEP **05** (2006), p. 026. arXiv: hep-ph/0603175.

[145] S. Frixione, E. Laenen, P. Motylinski, B. R. Webber and C. D. White. *Single-top hadroproduction in association with a W boson*. JHEP **07** (2008), p. 029. arXiv: 0805.3067 [hep-ph].

[146] T. Gleisberg and S. Hoche. *Comix, a new matrix element generator*. JHEP **12** (2008), p. 039. arXiv: 0808.3674 [hep-ph].

[147] F. Cascioli, P. Maierhofer and S. Pozzorini. *Scattering Amplitudes with Open Loops*. Phys. Rev. Lett. **108** (2012), p. 111601. arXiv: 1111.5206 [hep-ph].

[148] S. Schumann and F. Krauss. *A Parton shower algorithm based on Catani-Seymour dipole factorisation*. JHEP **03** (2008), p. 038. arXiv: 0709.1027 [hep-ph].

[149] ATLAS Collaboration. *Measurement of $W^{\pm}$ and Z Boson Production Cross Sections in pp Collisions at $\sqrt{s} = 13$ TeV with the ATLAS Detector*. ATLAS-CONF-2015-039. URL: https://cds.cern.ch/record/2045487.

[150] ATLAS Collaboration. *Multi-boson simulation for* 13 *TeV ATLAS analyses*. ATL-PHYS-PUB-2016-002. URL: https://cds.cern.ch/record/2119986.

[151] ATLAS Collaboration. *Estimation of non-prompt and fake lepton backgrounds in final states with top quarks produced in proton–proton collisions at* $\sqrt{s} = 8$ *TeV with the ATLAS Detector*. ATLAS-CONF-2014-058. URL: https://cds.cern.ch/record/1951336.

[152] J. Erdmann et al. *A likelihood-based reconstruction algorithm for top-quark pairs and the KLFitter framework*. Nucl. Instrum. Meth **748** (2014), pp. 18–25.

[153] ATLAS Collaboration. *Measurement of the top quark mass with the template method in the* $t\bar{t} \to$ *lepton + jets channel using ATLAS data*. Eur. Phys. J. C **72** (2012), p. 2046. arXiv: 1203.5755 [hep-ex].

[154] ATLAS Collaboration. *Direct top-quark decay width measurement in the* $t\bar{t}$ *lepton+jets channel at* $\sqrt{s} = 8$ *TeV with the ATLAS experiment*. ATLAS-CONF-2017-056. URL: https://cds.cern.ch/record/2273872.

[155] ATLAS Collaboration. *Luminosity determination in pp collisions at* $\sqrt{s} = 8$ *TeV using the ATLAS detector at the LHC*. Eur. Phys. J. C **76** (2016), p. 653. arXiv: 1608.03953 [hep-ex].

[156] ATLAS Collaboration. *Measurement of the inelastic proton-proton cross section at* $\sqrt{s} = 13$ *TeV with the ATLAS detector at the LHC*. Phys. Rev. Lett. **117** (2016), p. 182002. arXiv: 1606.02625 [hep-ex].

[157] J. W. Hetherly and S. Sekula. *Using VH Associated Production to Search for the* $b\bar{b}$ *Decay of the Higgs Boson with Data from the ATLAS Detector at* $\sqrt{s} =$ 13TeV. CERN-THESIS-2017-350. 2017. URL: https://cds.cern.ch/record/2313140.

[158] A. Bredenstein, A. Denner, S. Dittmaier and S. Pozzorini. *NLO QCD corrections to* $pp \to t\bar{t}b\bar{b} + X$ *at the LHC*. Phys. Rev. Lett. **103** (2009), p. 012002. arXiv: 0905.0110 [hep-ph].

[159] A. Bredenstein, A. Denner, S. Dittmaier and S. Pozzorini. *NLO QCD Corrections to Top Anti-Top Bottom Anti-Bottom Production at the LHC: 2. full hadronic results*. JHEP **03** (2010), p. 021. arXiv: 1001.4006 [hep-ph].

[160] G. Bevilacqua, M. Czakon, C. G. Papadopoulos and M. Worek. *Hadronic top-quark pair production in association with two jets at Next-to-Leading Order QCD*. Phys. Rev. D **84** (2011), p. 114017. arXiv: 1108.2851 [hep-ph].

[161] F. Cascioli, P. Maierhöfer, N. Moretti, S. Pozzorini and F. Siegert. *NLO matching for $t\bar{t}b\bar{b}$ production with massive b-quarks*. Phys. Lett. B **734** (2014), pp. 210–214. arXiv: 1309.5912 [hep-ph].

[162] ATLAS Collaboration. *Measurements of fiducial cross-sections for $t\bar{t}$ production with one or two additional b-jets in pp collisions at $\sqrt{s}$=8 TeV using the ATLAS detector*. Eur. Phys. J. C **76** (2016), p. 11.

[163] CMS Collaboration. *Measurements of $t\bar{t}$ cross sections in association with b-jets and inclusive jets and their ratio using dilepton final states in pp collisions at $\sqrt{s}$=13 TeV*. Phys. Lett. B **776** (2018), pp. 355–378.

[164] G. Bevilacqua and M. Worek. *On the ratio of $t\bar{t}b\bar{b}$ and $t\bar{t}jj$ cross sections at the CERN Large Hadron Collider*. JHEP **07** (2014), p. 135. arXiv: 1403.2046 [hep-ph].

[165] P. Artoisenet, R. Frederix, O. Mattelaer and R. Rietkerk. *Automatic spin-entangled decays of heavy resonances in Monte Carlo simulations*. JHEP **03** (2013), p. 015. arXiv: 1212.3460 [hep-ph].

[166] LHC Higgs Cross Section Working Group. *Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector* (2016). arXiv: 1610.07922 [hep-ph].

[167] R. Raitio and W. W. Wada. *Higgs-boson production at large transverse momentum in quantum chromodynamics*. Phys. Rev. D **19** (1979), pp. 941–944.

[168] W. Beenakker et al. *NLO QCD corrections to t anti-t H production in hadron collisions*. Nucl. Phys. B **653** (2003), pp. 151–203. arXiv: hep-ph/0211352 [hep-ph].

[169] S. Dawson, C. Jackson, L. H. Orr, L. Reina and D. Wackeroth. *Associated Higgs production with top quarks at the large hadron collider: NLO QCD corrections*. Phys. Rev. D **68** (2003), p. 034022. arXiv: hep-ph/0305087 [hep-ph].

[170] Y. Zhang, W.-G. Ma, R.-Y. Zhang, C. Chen and L. Guo. *QCD NLO and EW NLO corrections to $t\bar{t}H$ production with top quark decays at hadron collider*. Phys. Lett. B **738** (2014), pp. 1–5. arXiv: 1407.1110 [hep-ph].

[171] S. Frixione, V. Hirschi, D. Pagani, H.-S. Shao and M. Zaro. *Electroweak and QCD corrections to top-pair hadroproduction in association with heavy bosons*. JHEP **06** (2015), p. 184. arXiv: 1504.03446 [hep-ph].

[172] ATLAS Collaboration. *Search for the standard model Higgs boson produced in association with top quarks and decaying into a $b\bar{b}$ pair in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*. Phys. Rev. D **97** (2018), p. 072016. arXiv: 1712.08895 [hep-ex].

[173] B. Nachman, P. Nef, A. Schwartzman, M. Swiatlowski and C. Wanotayaroj. *Jets from jets: re-clustering as a tool for large radius jet reconstruction and grooming at the LHC*. JHEP **02** (2015), p. 075. arXiv: 1407.2922 [hep-ph].

[174] ATLAS Collaboration. *Search for the Standard Model Higgs boson produced in association with top quarks and decaying into $b\bar{b}$ in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector*. Eur. Phys. J. C **75** (2015), p. 349. arXiv: 1503.05066 [hep-ex].

[175] G. Cowan, K. Cranmer, E. Gross and O. Vitells. *Asymptotic formulae for likelihood-based tests of new physics*. Eur. Phys. J. C **71** (2011). [Erratum: Eur. Phys. J. C 73 (2013)], p. 1554.

[176] T. Junk. *Confidence level computation for combining searches with small statistics*. Nucl. Instrum. Methods Phys. Res., Section A **434** (1999), pp. 435–443.

[177] F. James. *Statistical methods in experimental physics*. Ed. by World Scientific. 2006. ISBN: 9789812567956.

[178] A. Wald. *Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large*. Trans. Amer. Math. Soc **54** (1943), pp. 426–482.

[179] S. S. Wilks. *The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses*. Ann. Math. Statist. **9** (1938), pp. 60–62.

[180] L. Lyons. *Discovering the Significance of 5 sigma* (2013). arXiv: 1310.1284v1 [physics.data-an].

[181] A. L. Read. *Presentation of search results: the CLs technique*. J. Phys. G **28** (2002), p. 2693.

[182] ATLAS Collaboration. *Studies of $t\bar{t}+c\bar{c}$ production with MadGraph5_aMC@NLO and Herwig++ for the ATLAS experiment*. ATL-PHYS-PUB-2016-011. 2016. URL: https://cds.cern.ch/record/2153876.

[183] J. M. Campbell and R. K. Ellis. *$t\bar{t}W^{\pm}$ production and decay at NLO*. JHEP **07** (2012), p. 052. arXiv: 1204.5678 [hep-ph].

[184] ATLAS Collaboration. *Evidence for the associated production of the Higgs boson and a top quark pair with the ATLAS detector*. Phys. Rev. D **97** (2018), p. 072003.

[185] ATLAS Collaboration. *Measurement of the Higgs boson coupling properties in the $H \to ZZ^* \to 4\ell$ decay channel at $\sqrt{s} = 13$ TeV with the ATLAS detector*. JHEP **03** (2018), p. 95. arXiv: 1712.02304 [hep-ex].

[186] ATLAS Collaboration. *Measurements of Higgs boson properties in the diphoton decay channel with 36 fb^{-1} of pp collision data at $\sqrt{s} = 13$ TeV with the ATLAS detector* (2018). arXiv: 1802.04146 [hep-ex].

[187] ATLAS and CMS Collaborations. *Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysis of the LHC pp collision data at $\sqrt{s}$ = 7 and 8 TeV*. JHEP **2016** (2016), p. 45.

[188] ATLAS Collaboration. *Observation of Higgs boson production in association with a top quark pair at the LHC with the ATLAS detector*. Phys. Lett. B **784** (2018), pp. 173–191.

[189] ATLAS Collaboration. *Projections for measurements of Higgs boson signal strengths and coupling parameters with the ATLAS detector at the HL-LHC*. ATL-PHYS-PUB-2014-016. URL: https://cds.cern.ch/record/1956710.

[190] ATLAS Collaboration. *Jet mass and substructure of inclusive jets in $\sqrt{s} = 7$ TeV pp collisions with the ATLAS experiment*. JHEP **05** (2012), p. 128. arXiv: 1203.4606 [hep-ex].

[191] V. Barger, J. Ohnemus and R. Phillips. *Event shape criteria for single lepton top signals*. Phys. Rev. D **48** (1993), p. 3953. arXiv: `hep-ph/9308216`.

RecoBDT variables

Table A.1: Input variables for the reconstruction BDTs used in the ICHEP analysis. The subscript had (lep) indicates the hadronically (leptonically) decaying W or t and q_i refers to quarks from W. RecoBDT trained without information from the Higgs exploits only topological information from $t\bar{t}$.

Variable	($\geq$ 6j, $\geq$ 4b)	($\geq$ 6j, 3b)	(5j, $\geq$ 4b)
Topological information from $t\bar{t}$			
t_{lep} mass	✓	✓	✓
t_{had} mass	✓	✓	–
Incomplete t_{had} mass	–	–	✓
W_{had} mass	✓	✓	–
Mass of W_{had} and b from t_{lep}	✓	✓	–
Mass of q from W_{had} and b from t_{lep}	–	–	✓
Mass of W_{lep} and b from t_{had}	✓	✓	✓
$\Delta R(W_{\text{had}}, b \text{ from } t_{\text{had}})$	✓	✓	–
$\Delta R(q \text{ from } W_{\text{had}}, b \text{ from } t_{\text{had}})$	–	–	✓
$\Delta R(W_{\text{had}}, b \text{ from } t_{\text{lep}})$	✓	✓	–
$\Delta R(q \text{ from } W_{\text{had}}, b \text{ from } t_{\text{lep}})$	–	–	✓
$\Delta R(\text{lep}, b \text{ from } t_{\text{lep}})$	✓	✓	✓
$\Delta R(\text{lep}, b \text{ from } t_{\text{had}})$	✓	✓	✓
$\Delta R(b \text{ from } t_{\text{lep}}, b \text{ from } t_{\text{had}})$	✓	✓	✓
$\Delta R(q_1 \text{ from } W_{\text{had}}, q_2 \text{ from } W_{\text{had}})$	✓	✓	–
$\Delta R(b \text{ from } t_{\text{had}}, q_1 \text{ from } W_{\text{had}})$	✓	✓	–
$\Delta R(b \text{ from } t_{\text{had}}, q_2 \text{ from } W_{\text{had}})$	✓	✓	–
min. $\Delta R(b \text{ from } t_{\text{had}}, q \text{ from } W_{\text{had}})$	✓	✓	–
$\Delta R(\text{lep}, b \text{ from } t_{\text{lep}})$ - min. $\Delta R(b \text{ from } t_{\text{had}}, q \text{ from } W_{\text{had}})$	✓	✓	–
$\Delta R(\text{lep}, b \text{ from } t_{\text{lep}})$ - $\Delta R(b \text{ from } t_{\text{had}}, q \text{ from } W_{\text{had}})$	–	–	✓
Topological information from Higgs			
Higgs mass	✓	✓	✓
Mass of Higgs and q_1 from W_{had}	✓	✓	✓
$\Delta R(b_1 \text{ from Higgs}, b_2 \text{ from Higgs})$	✓	✓	✓
$\Delta R(b_1 \text{ from Higgs}, \text{lep})$	✓	✓	✓
$\Delta R(b_1 \text{ from Higgs}, b \text{ from } t_{\text{lep}})$	–	✓	✓
$\Delta R(b_1 \text{ from Higgs}, b \text{ from } t_{\text{had}})$	–	✓	✓

RecoJVC

B

B.1 Input variables

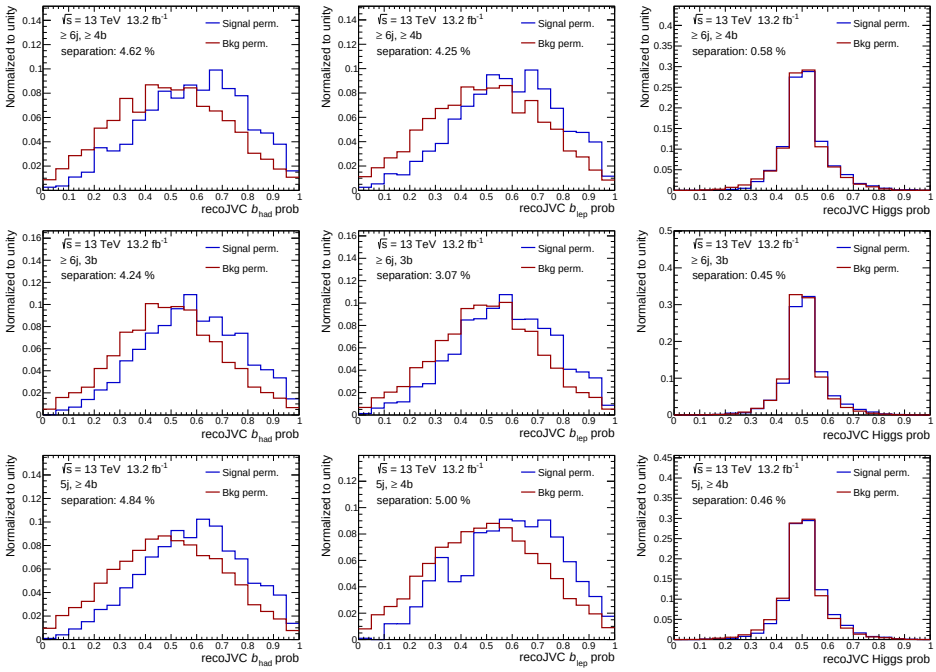


Figure B.1: Distribution for the b_{had} (left), b_{lep} (centre) and Higgs (right) probability as defined by Equations 5.1 and 5.2 for the signal and background permutation for the three signal regions of the ICHEP analysis. The values of the separation for each variable in each region is reported.

B.2 Output variables of recoJVC

($\geq$ 6j, $\geq$ 4b) signal region

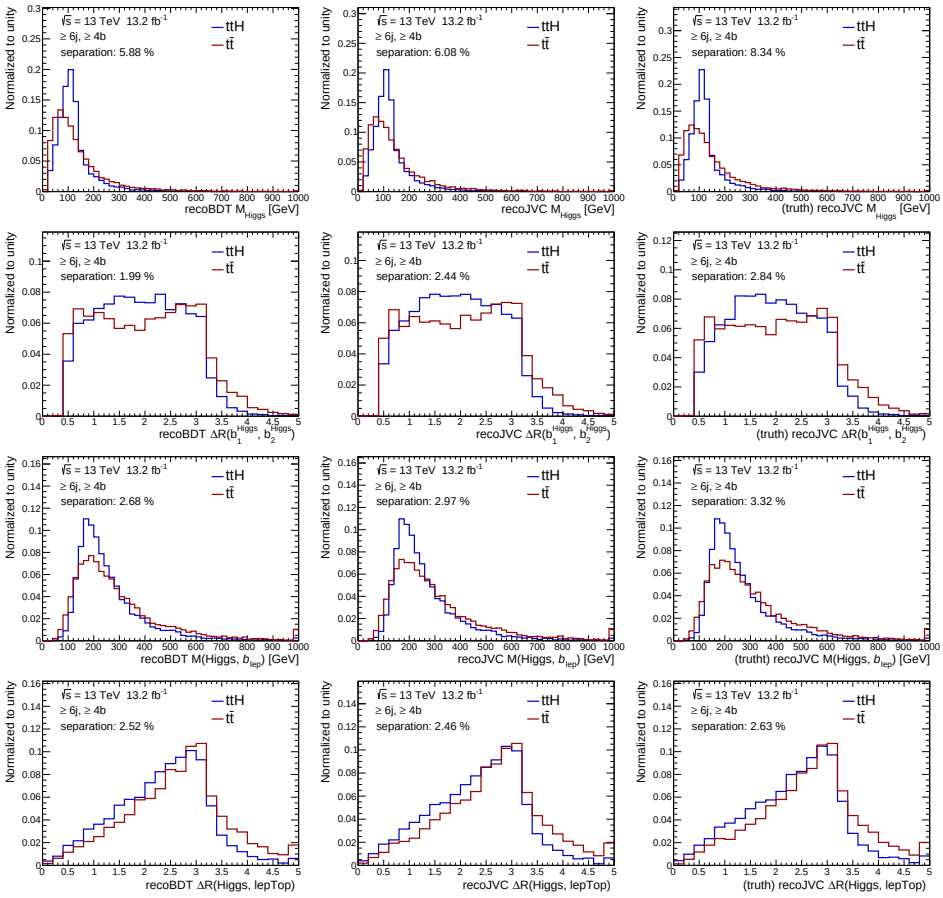


Figure B.2: Comparison of the $t\bar{t}H(b\bar{b})$ and $t\bar{t}$ distributions of the Higgs candidate mass (first row), ΔR(b1Higgs, b2Higgs) (second row), invariant mass of the Higgs and b_{lep} (third row) and ΔR(Higgs, lepTop) (last row) in the ($\geq$ 6j, $\geq$ 4b) SR for the three training configurations of the recoJVC: default training (left), with charge-variables (centre) and with truth charge (right).

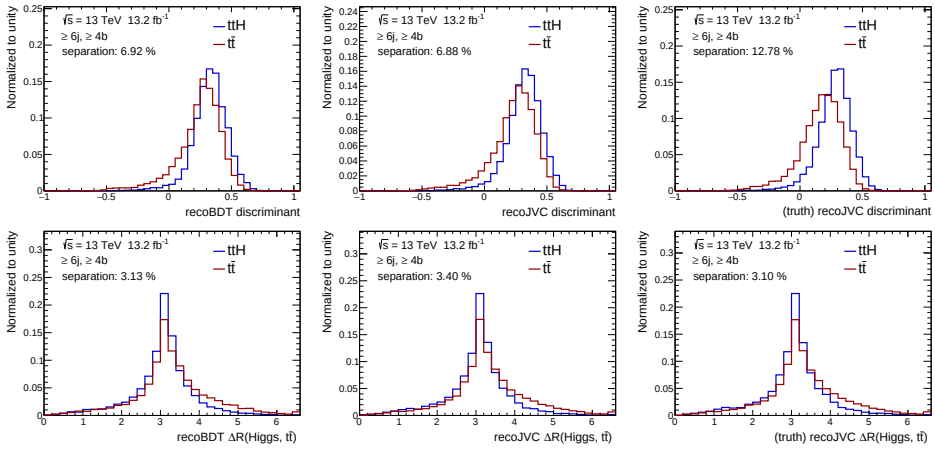


Figure B.3: Comparison of the $t\bar{t}H(b\bar{b})$ and $t\bar{t}$ distributions of the recoJVC output discriminant (top) and ΔR(Higgs, $t\bar{t}$) (bottom) in the ($\geq$ 6j, $\geq$ 4b) SR for the three training configurations of the recoJVC: default training (left), with charge-variables (centre) and with truth charge (right).

($\geq$ 6j, 3b) **signal region**

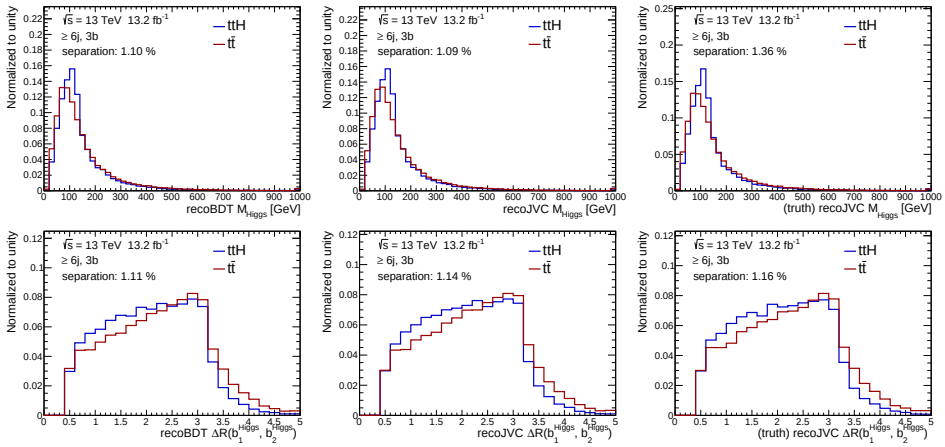


Figure B.4: Comparison of the $t\bar{t}H(b\bar{b})$ and $t\bar{t}$ distributions of the Higgs candidate mass (top) and ΔR(b1Higgs, b2Higgs) (bottom) in the ($\geq$ 6j, 3b) SR for the three training configurations of the recoJVC: default training (left), with charge-variables (centre) and with truth charge (right).

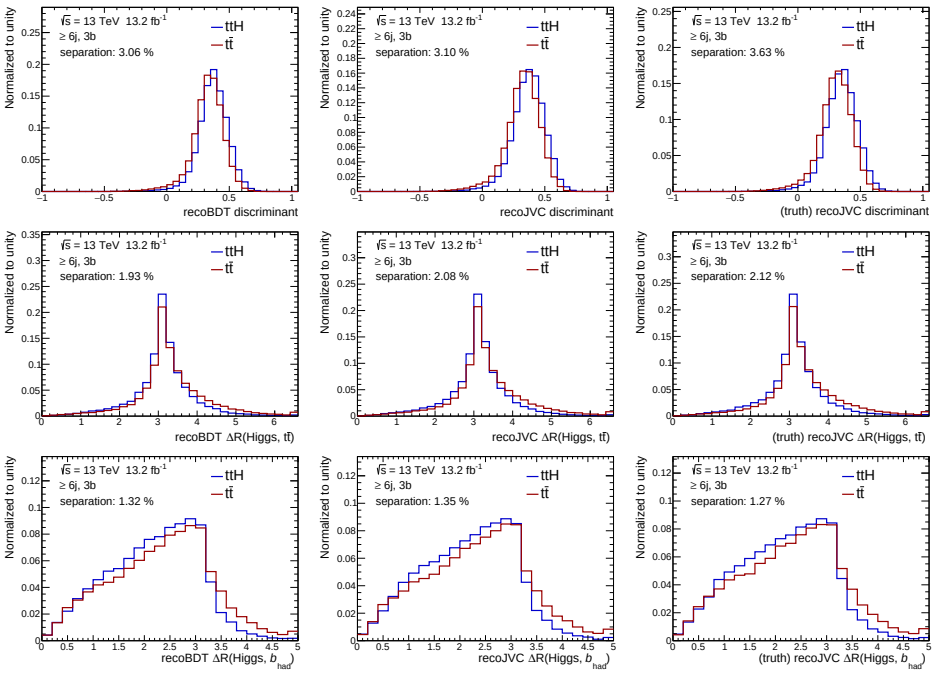


Figure B.5: Comparison of the $t\bar{t}H(b\bar{b})$ and $t\bar{t}$ distributions of the recoJVC output discriminant (top), ΔR(Higgs, $t\bar{t}$) (middle) and ΔR(Higgs, bhadtop) (bottom) in the ($\geq$ 6j, 3b) SR for the three training configurations of the recoJVC: default training (left), with charge-variables (centre) and with truth charge (right).

$(5j, \geq 4b)$ **signal region**

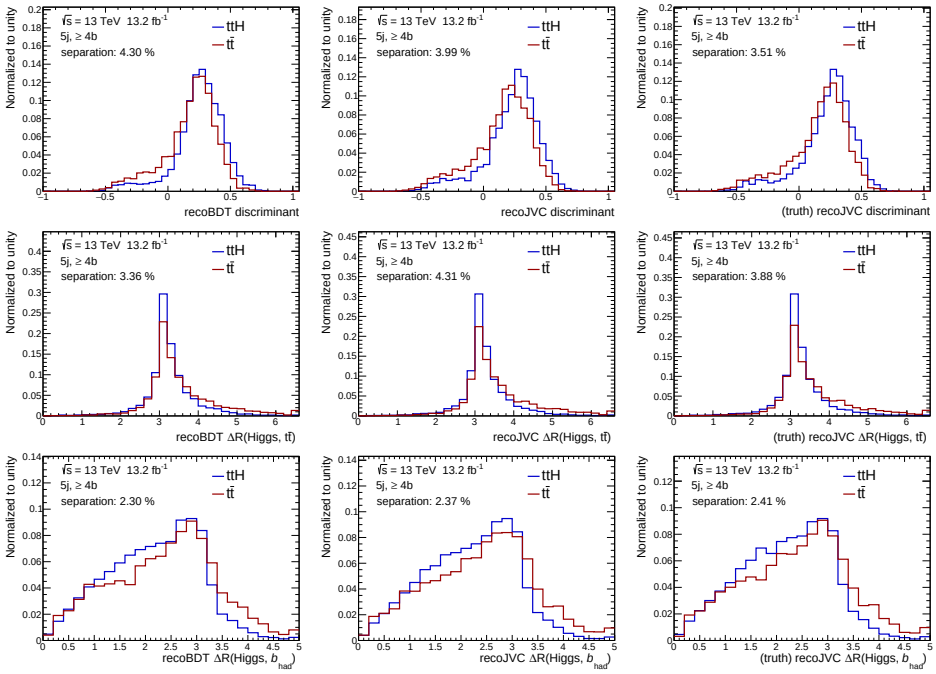


Figure B.6: Comparison of the $t\bar{t}H(b\bar{b})$ and $t\bar{t}$ distributions of the recoJVC output discriminant (top), ΔR(Higgs, $t\bar{t}$) (middle) and ΔR(Higgs, bhadtop) (bottom) in the $(5j, \geq 4b)$ SR for the three training configurations of the recoJVC: default training (left), with charge-variables (centre) and with truth charge (right).

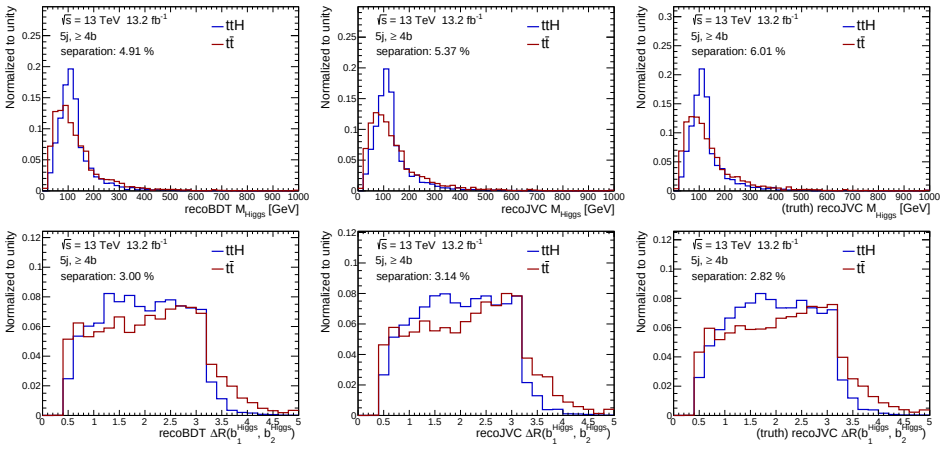


Figure B.7: Comparison of the $t\bar{t}H(b\bar{b})$ and $t\bar{t}$ distributions of the Higgs candidate mass (top) and ΔR(b1Higgs, b2Higgs) (bottom) in the (5j, $\geq$ 4b) SR for the three training configurations of the recoJVC: default training (left), with charge-variables (centre) and with truth charge (right).

ClassBDT inputs for the ICHEP analysis

C

Table C.1: List of the input variables for the classification BDT used in the ICHEP analysis.

	($\geq$ 6j, $\geq$ 4b)	($\geq$ 6j, 3b)	(5j, $\geq$ 4b)
Kinematic variables			
Centrality	✓	✓	✓
$\Delta\eta_{jj}^{\text{max }\Delta\eta}$	✓	✓	✓
$H1$	✓	✓	✓
p_{T}^{jet5}	✓	✓	✓
$\Delta R_{bb}^{\text{avg}}$	✓	✓	✓
Aplan	✓	✓	✓
N_{30}^{Higgs}	✓	–	✓
$m_{bb}^{\text{min }\Delta R}$	✓	✓	–
$m_{bj}^{\text{max }p_{T}}$	–	✓	–
$\Delta R_{bb}^{\text{max }p_{T}}$	✓	–	–
$\Delta R_{\text{lep}-bb}^{\text{min }\Delta R}$	–	–	✓
N_{40}^{jet}	–	✓	–
H_{T}^{had}	–	✓	✓
$m_{jj}^{\text{min }\Delta R}$	–	–	✓
Variables from recoBDT			
Higgs mass	✓	✓	✓
ΔR(b1Higgs, b2Higgs)	✓	✓	✓
(Higgs+blepTop)_mass	✓	–	–
ΔR(Higgs, lepTop)	✓	–	–
Variables from recoBDT_withHiggs			
BDT output	✓	✓	✓
ΔR(Higgs, $t\bar{t}$)	✓	✓	✓
ΔR(Higgs, bhadtop)	–	✓	✓

HFBDT

D

D.1 Input variables separation

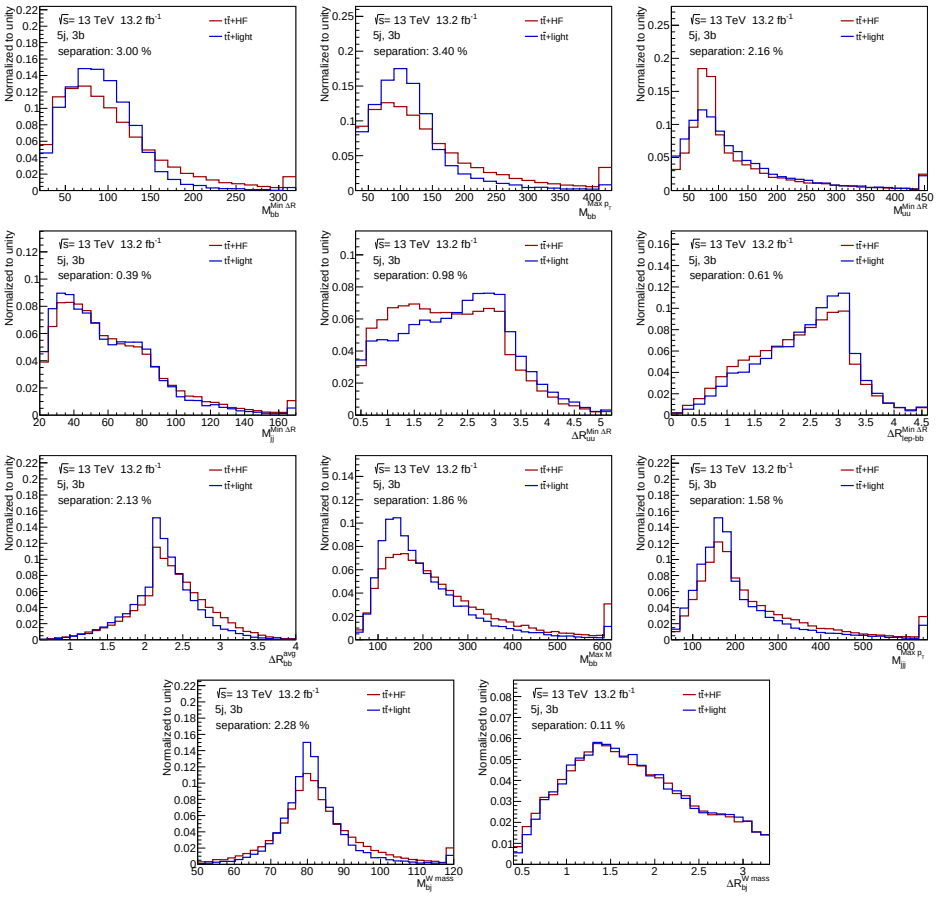


Figure D.1: Separation between the $t\bar{t}$+HF and $t\bar{t}+light$ processes for the input variables of the HFBDT.

D.2 Input variables correlation

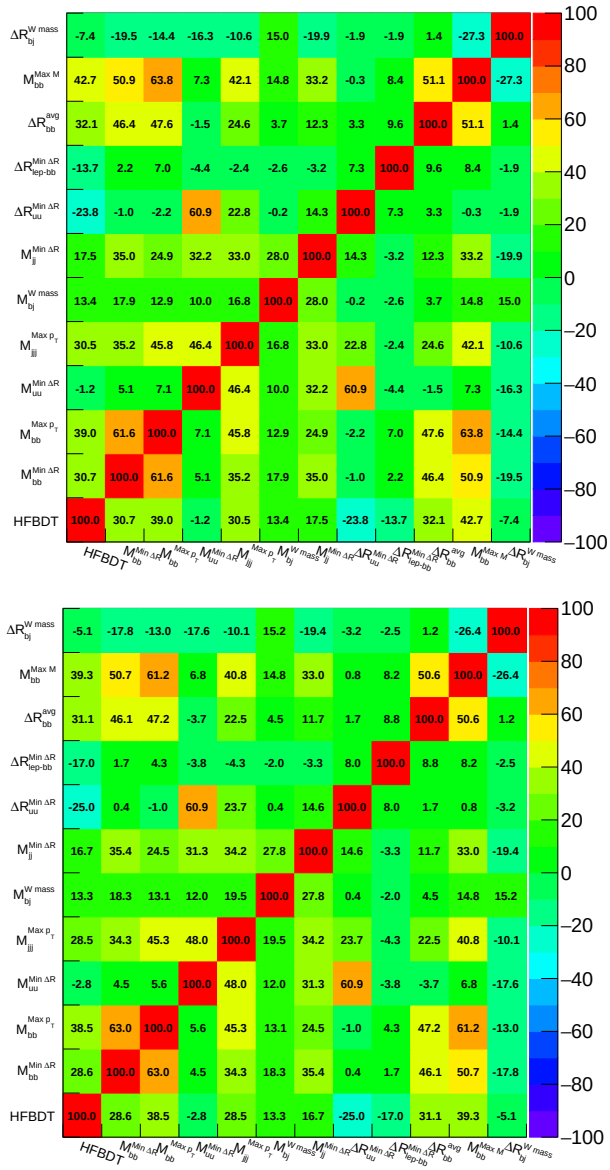


Figure D.2: Linear correlations of the input variables and the final discriminant for data (top) and MC (bottom) for the HFBDT.

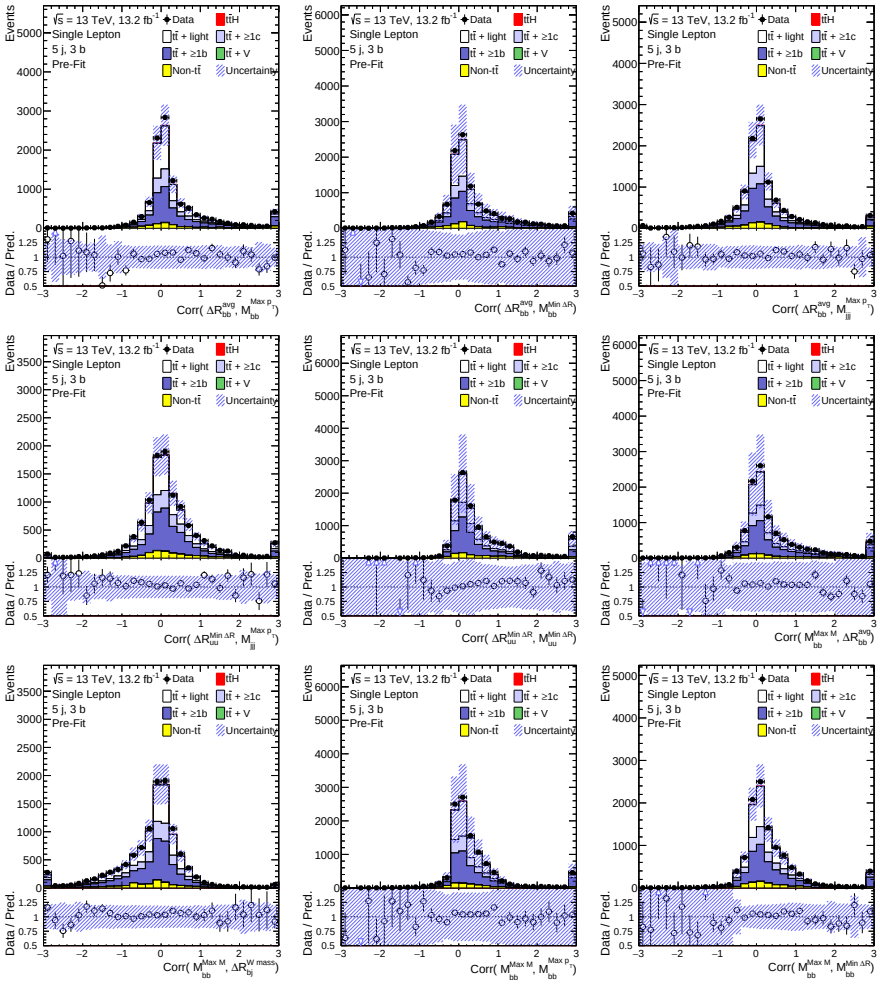


Figure D.3: First set of pre-fit linear correlations between the input variables for the HFBDT.

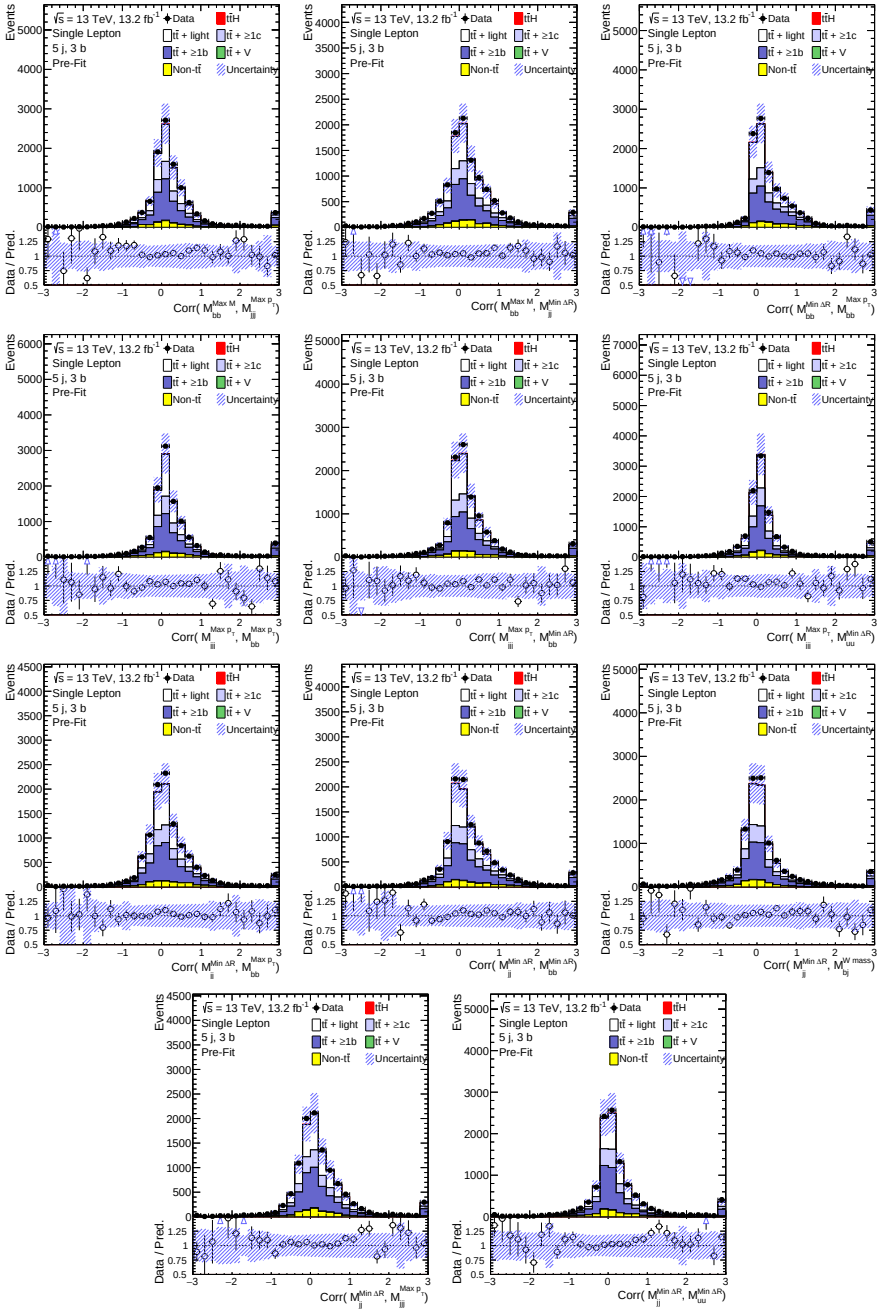


Figure D.4: Second set of pre-fit linear correlations between the input variables for the HFBDT.

ClassBDT inputs for the paper analysis

E

The full list of variables used as inputs to the classification BDTs in each of the signal regions is reported.

Table E.1: Input variables to the classification BDT in the boosted single-lepton signal region. Additional b-jets are b-jets not contained in the Higgs-boson and top-quark candidates.

Variable	Definition
Variables from jet reclustering	
$\Delta R_{H,t}$	ΔR between the Higgs-boson and top-quark candidates
$\Delta R_{t,b^{\text{add}}}$	ΔR between the top-quark candidate and additional b-jet
$\Delta R_{H,b^{\text{add}}}$	ΔR between the Higgs-boson candidate and additional b-jet
$\Delta R_{H,\ell}$	ΔR between the Higgs-boson candidate and lepton
$m_{\text{Higgs candidate}}$	Higgs-boson candidate mass
$\sqrt{d_{12}}$	Top-quark candidate first splitting scale [190]
Variables from b-tagging	
$w_{b\text{-tag}}$	Sum of b-tagging discriminants of all b-jets
$w^{\text{add}}_{b\text{-tag}}/w_{b\text{-tag}}$	Ratio of sum of b-tagging discriminants of additional b-jets to all b-jets

Table E.2: Input variables to the classification BDTs in the single-lepton signal regions. For variables from the reconstruction BDT, those with a * are from the BDT using Higgs-boson information, those with no * are from the BDT without Higgs-boson information.

Variable	Definition	$SR_{1,2,3}^{\geq 6j}$	$SR_{1,2}^{5j}$
General kinematic variables			
ΔR_{bb}^{avg}	Average ΔR for all b-tagged jet pairs	✓	✓
$\Delta R_{bb}^{max\ p_T}$	ΔR between the two b-tagged jets with the largest vector sum p_T	✓	–
$\Delta\eta_{jj}^{max\ \Delta\eta}$	Maximum $\Delta\eta$ between any two jets	✓	✓
$m_{bb}^{min\ \Delta R}$	Mass of the combination of two b-tagged jets with the smallest ΔR	✓	–
$m_{jj}^{min\ \Delta R}$	Mass of the combination of any two jets with the smallest ΔR	–	✓
$N_{bb}^{Higgs\ 30}$	Number of b-tagged jet pairs with invariant mass within 30 GeV of the Higgs-boson mass	✓	✓
H_T^{had}	Scalar sum of jet p_T	–	✓
$\Delta R_{lep-bb}^{min\ \Delta R}$	ΔR between the lepton and the combination of the two b-tagged jets with the smallest ΔR	–	✓
Aplan	$1.5\lambda_2$, where λ_2 is the second eigenvalue of the momentum tensor [191] built with all jets	✓	✓
H_1	Second Fox–Wolfram moment computed using all jets and the lepton	✓	✓
Variables from reconstruction BDT			
BDT output	Output of the reconstruction BDT	✓*	✓*
m_{bb}^{Higgs}	Higgs candidate mass	✓	✓
$m_{H,b_{lep\ top}}$	Mass of Higgs candidate and b-jet from leptonic top candidate	✓	–
ΔR_{bb}^{Higgs}	ΔR between b-jets from the Higgs candidate	✓	✓
$\Delta R_{H,t\bar{t}}$	ΔR between Higgs candidate and $t\bar{t}$ candidate system	✓*	✓*
$\Delta R_{H,lep\ top}$	ΔR between Higgs candidate and leptonic top candidate	✓	–
$\Delta R_{H,b_{had\ top}}$	ΔR between Higgs candidate and b-jet from hadronic top candidate	–	✓*
Variables from likelihood and matrix element method calculations			
LHD	Likelihood discriminant	✓	✓
MEM_{D1}	Matrix element discriminant (in $SR_1^{\geq 6j}$ only)	✓	–
Variables from b-tagging (not in $SR_1^{\geq 6j}$)			
$w_{b\text{-tag}}^{Higgs}$	Sum of b-tagging discriminants of jets from best Higgs candidate from the reconstruction BDT	✓	✓
B_{jet}^3	3rd largest jet b-tagging discriminant	✓	✓
B_{jet}^4	4th largest jet b-tagging discriminant	✓	✓
B_{jet}^5	5th largest jet b-tagging discriminant	✓	✓

Table E.3: Variables used in the classification BDTs in the dilepton signal regions. For variables from the reconstruction BDT, those with a * are from the BDT using Higgs-boson information, those with no * are from the BDT without Higgs-boson information while for those with a ** both versions are used.

Variable	Definition	$SR_1^{\geq 4j}$	$SR_2^{\geq 4j}$	$SR_3^{\geq 4j}$
General kinematic variables				
m_{bb}^{min}	Minimum invariant mass of a b-tagged jet pair	✓	✓	–
$m_{\text{bb}}^{\text{max M}}$	Maximum invariant mass of a b-tagged jet pair	–	–	✓
$m_{\text{bb}}^{\text{min }\Delta R}$	Invariant mass of the b-tagged jet pair with minimum ΔR	✓	–	✓
$m_{\text{jj}}^{\text{max }p_{\text{T}}}$	Invariant mass of the jet pair with maximum p_{T}	✓	–	–
$m_{\text{bb}}^{\text{max }p_{\text{T}}}$	Invariant mass of the b-tagged jet pair with maximum p_{T}	✓	–	✓
$\Delta\eta_{bb}^{\text{avg}}$	Average $\Delta\eta$ for all b-tagged jet pairs	✓	✓	✓
$\Delta\eta_{\ell,\text{j}}^{\text{max}}$	Maximum $\Delta\eta$ between a jet and a lepton	–	✓	✓
$\Delta R_{\text{bb}}^{\text{max }p_{\text{T}}}$	ΔR between the b-tagged jet pair with maximum p_{T}	–	✓	✓
$N_{bb}^{\text{Higgs 30}}$	Number of b-tagged jet pairs with invariant mass within 30 GeV of the Higgs-boson mass	✓	✓	–
$n_{\text{jets}}^{p_{\text{T}}>40}$	Number of jets with $p_{\text{T}} > 40$ GeV	–	✓	✓
Aplan_{b-jet}	$1.5\lambda_2$, where λ_2 is the second eigenvalue of the momentum tensor [191] built with all b-tagged jets	–	✓	–
$H_{\text{T}}^{\text{all}}$	Scalar sum of p_{T} of all jets and leptons	–	–	✓
Variables from reconstruction BDT				
BDT output	Output of the reconstruction BDT	✓**	✓**	✓
m_{bb}^{Higgs}	Higgs candidate mass	✓	–	✓
$\Delta R_{H,t\bar{t}}$	ΔR between Higgs candidate and $t\bar{t}$ candidate system	✓*	–	–
$\Delta R_{H,\ell}^{\text{min}}$	Minimum ΔR between Higgs candidate and lepton	✓	✓	✓
$\Delta R_{H,b}^{\text{min}}$	Minimum ΔR between Higgs candidate and b-jet from top	✓	✓	–
$\Delta R_{H,b}^{\text{max}}$	Maximum ΔR between Higgs candidate and b-jet from top	–	✓	–
$\Delta R_{bb}^{\text{Higgs}}$	ΔR between the two jets matched to the Higgs candidate	–	✓	–
Variables from b-tagging				
$w_{b\text{-tag}}^{\text{Higgs}}$	Sum of b-tagging discriminants of jets from best Higgs candidate from the reconstruction BDT	–	✓	–

Summary

The Standard Model (SM) is the theory that best summarizes the knowledge, as of today, of the subatomic world. The fundamental building blocks of matter are half-integer spin particles, called fermions. Three out of the four fundamental forces are described by the SM: the electromagnetic, weak and strong interactions are rendered with the exchange of force carriers particles with integer spin, called bosons. Gravity is not included into the theory.

There are a total of six leptons and six quarks organized in three families or generations, but only the first generation is needed to build up protons, neutrons, atoms and molecules. The electromagnetic force is mediated by one of the most known particles, the photon γ; the weak force, responsible for radioactive decays, is mediated by the $W^{\pm}$ and Z bosons; and the strong force, responsible for confining quarks into hadrons, is mediated by gluons g.

Gauge bosons in the SM must be massless for it to be self-consistent, a fact that is clearly against experimental evidence. In order to solve this problem, the Brout-Englert-Higgs mechanism comes to rescue: the Higgs field spontaneously acquires a non-vanishing vacuum expectation value, thus breaking the original electroweak symmetry, and the interaction between the gauge bosons and the Higgs field gives rise to their masses. The same mechanism can be used to give mass to the fermions via the so-called Yukawa interaction.

The Brout-Englert-Higgs mechanism predicts the existence of a particle with unknown mass: the famous Higgs boson. A particle with a mass near 125 GeV, which resembles closely the Higgs boson, was discovered by the ATLAS and CMS Collaborations in 2012, providing experimental proof to the correctness of this mechanism.

Once its mass is established, all properties, such as the couplings to

the other particles, can be predicted. Up to today, no significant deviation from the the predicted values is observed. Nevertheless, even if the SM proved to be a very successful theory, it cannot be the ultimate theory of Nature, as it has many open fundamental questions, such as the nature of dark matter or the hierarchy problem. Many alternative theories that address these issues predict different properties for the Higgs bosons (or the existence of more than one). On the other hand, given that no sign of new physics is visible at the horizon, precise measurements in the Higgs sector are therefore of high priority for the particle physics community, in order to test the internal consistency of the SM.

In this respect, measuring the Higgs couplings to the last generation of quarks is of great importance. Indirect measurements of the top Yukawa coupling have been performed by inferring the value from the gluon fusion process, however other particles can enter in such loops (even beyond the SM ones). Alternatively, it is possible to measure directly this coupling via the $t\bar{t}H$ production mode.

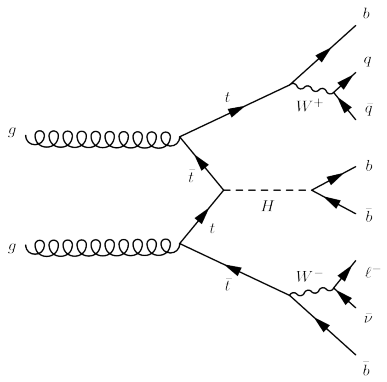


Figure S.1: Feynman diagram for the $t\bar{t}H(b\bar{b})$ process.

This thesis presents the search for the Standard Model Higgs boson produced in association with a pair of top quarks and decaying into a $b\bar{b}$ pair, $t\bar{t}H(b\bar{b})$. The top quark decays (almost) exclusively into a real W boson and b-quark. The $t\bar{t}$ decay topology determines the analysis channels effectively used: dilepton and (resolved and boosted) single-lepton. One Feynman diagram illustrating the process is shown in Figure S.1.

This analysis presents various difficulties, all traceable to the large irreducible $t\bar{t}$ + jets background. Notably, in the selected phase space of the analysis, the extra jets produced in association with the $t\bar{t}$ pair are often heavy flavoured jets, i.e. jets originating from b- or c-quark fragmentation. These processes are not only difficult to model, but also present large theoretical uncertainties.

The serious problems posed by the overwhelming $t\bar{t} + \geq 1b$ background requires a complex analysis strategy employing several multivariate techniques. Selected events are first categorized into signal- and background-enriched regions. In the signal regions various methods are employed to reconstruct the final state; subsequently, Boosted Decision Trees (BDTs) are used to classify the events into signal- and background-like events. Finally, all regions are used as input to a profiled likelihood fit to data to determine simultaneously the event yields for the physics processes considered and to constrain the overall background model within the assigned systematic uncertainties.

Given the many jets in the final state, the large combinatorics complicates the identification of the Higgs boson decay products and the reconstruction the Higgs mass peak. The Jet Vertex Charge tagger (JVC) was developed specifically to improve this aspect.

It rests on the fact that the jet-to-quark assignment can benefit from the measurement of the electric charge of the b-quark that triggered the jet formation, therefore facilitating the reconstruction of the final state, the identification of the Higgs decay products and, consequently, increasing the analysis sensitivity. Both the algorithm and its calibration analysis are described in Chapter 4.

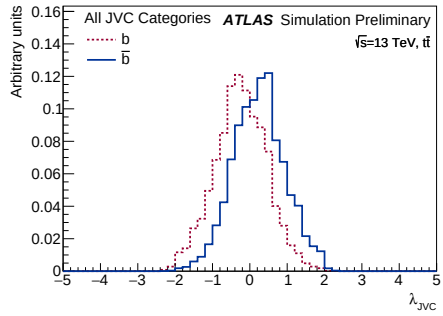


Figure S.2: λ_{JVC} distributions normalized to unity for positive (solid blue line) and negative (dashed red line) truth b-jets.

JVC uses the distinctive signs of the b-hadron decay chain to construct several variables sensitive to the quark charge that are combined by means of a Neural Network. Eventually, the likelihood ratio of the hypotheses for a b-jet to have a positive or negative charge is used as final discriminant, λ_{JVC}, whose distribution is shown in Figure S.2. Therefore, JVC allows to tag a jet as coming from a positively or negatively charged b-hadron.

The Jet Vertex Charge implementation into the $t\bar{t}H(b\bar{b})$ analysis is

presented in the first part of Chapter 5, for what is referred to as the ICHEP analysis. In spite of the improvement in reconstructing the final state, only a marginal improvement in discriminating the $t\bar{t}H(b\bar{b})$ signal from the $t\bar{t}+\geq 1b$ main background is observed, due to the fact that the $t\bar{t}+\geq 1b$ background has the same charge signature as the signal $t\bar{t}H(b\bar{b})$. For this reason, it was not used in the final analysis.

In order to improve the knowledge of the $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ backgrounds, the HFBDT was developed for the ICHEP analysis. A BDT is trained to disentangle the heavy flavour from the $t\bar{t}+light$ component, in order for the profiled likelihood fit to have a better handling on their systematic uncertainties. However, no significant improvement was observed, thus it was not employed in the final analysis design.

The second part of Chapter 5 contains the description of the so-called "paper" analysis. With respect to the ICHEP analysis, it is an updated version that contains several refinements and improvements, most notably a refined definition of signal and background regions. The results are based on the proton-proton collision data collected with the ATLAS detector at $\sqrt{s}$ = 13 TeV during 2015 and 2016, for a total of 36.1 fb^{-1} of integrated luminosity.

The best-fit value of the measured-to-predicted cross sections ratio, the signal strength μ, in all the single-lepton and dilepton regions yields a value of:

$$\mu = \frac{\sigma_{\text{meas}}}{\sigma_{\text{SM}}} = 0.84 \pm 0.29\ (\text{stat.})\ ^{+0.57}_{-0.54}\ (\text{syst.}) = 0.84^{+0.64}_{-0.61}$$

with the expected uncertainty of the signal strength identical to the measured one. It corresponds to an excess of 1.4 standard deviations over the expected SM background.

Given that the $t\bar{t}H$ cross section accounts for about 1% of the total Higgs boson cross section, in order to bring additional sensitivity to the $t\bar{t}H$ production, complementary analyses exploiting different decay modes have been performed: the Higgs boson decaying into a pair of photons, $t\bar{t}H(\rightarrow \gamma\gamma)$; the Higgs boson decay into four leptons, $t\bar{t}H(\rightarrow ZZ^* \rightarrow 4\ell)$; and the decay into a multi-lepton final state, $t\bar{t}H(\rightarrow WW^*/\tau\tau/ZZ^* \rightarrow \text{leptons})$, referred to as $H \rightarrow$ ML.

These analyses are performed on 36.1 fb^{-1} of integrated luminosity as well. A combined likelihood, which allows to properly account for common systematic uncertainties, is obtained from the product of individual likelihoods of each analysis. The best-fit value of the $t\bar{t}H$ signal strength determined from the combined likelihood equals:

$$\mu = 1.17 \pm 0.19\ (\text{stat.})\ ^{+0.27}_{-0.23}\ (\text{syst.})$$

which corresponds to an excess of events over the expected SM background with an observed (expected) significance of 4.2 (3.8) σ: it represents the first evidence for the $t\bar{t}H$ production mechanism found by ATLAS alone. The cross section for $t\bar{t}H$ production corresponding to the best-fit value of μ is 590^{+160}_{-150} fb, well compatible with the SM prediction of $\sigma^{\text{SM}}_{t\bar{t}H} = 507^{+35}_{-50}$ fb.

The observed signal strengths for the individual analyses and their combination are shown in Figure S.3a. Given that the four analyses have different acceptances in their analysis regions for the various Higgs boson decay modes, it is possible to independently determine μ for different Higgs decays. Figure S.3b shows the result of a fit to four signal strengths, one for each of the decay $H \to \tau\tau$, $H \to \gamma\gamma$, $H \to b\bar{b}$ and $H \to WW/ZZ$. All measurements are in agreement with the SM predictions and no significant deviation is observed.

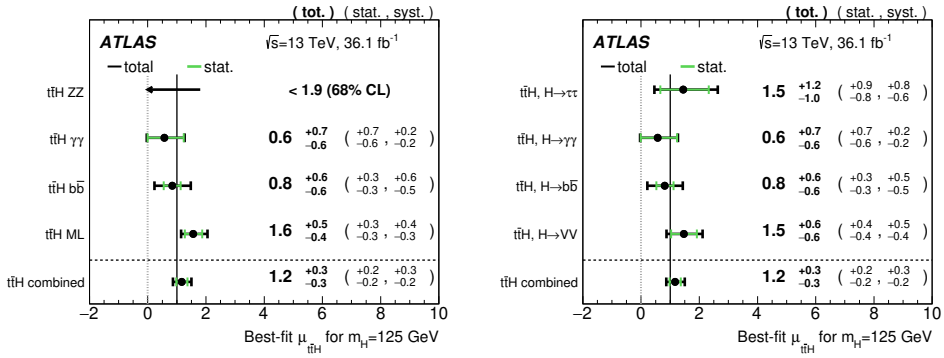


Figure S.3: Summary of the measurements of μ from individual analyses and their combination (left) and summary of the best-fit values of μ broken down by Higgs boson decay mode (right). As no events are observed in the $H \to 4\ell$ analysis, a 68% confidence level upper limit on μ is reported. The category $H \to VV$ combines both $H \to WW^*$ and $H \to ZZ^*$.

Samenvatting

Het standaardmodel (SM) is de theorie die de hedendaagse kennis van de subatomaire wereld beschrijft. De fundamentele bouwstenen van materie zijn deeltjes met een halftallige spin, zogeheten fermionen. Drie van de vier fundamentele krachten worden beschreven door het SM: de elektromagnetische kracht, de zwakke kernkracht, en de sterke kernkracht worden overgebracht door boodschapperdeeltjes met een heeltallige spin: de bosonen. Zwaartekracht wordt niet beschreven binnen het Standaard Model.

In totaal zijn er zes leptonen and zes quarks, onderverdeeld in drie generaties, echter de eerste generatie bevat alle bouwstenen voor protonen, neutronen, atomen en moleculen. De elektromagnetische kracht wordt overgebracht door een welbekend deeltje: het foton γ; de zwakke kernkracht, verantwoordelijk voor radioactief verval, wordt overgedragen door de $W^{\pm}$ en Z bosonen; en de sterke kernkracht, dat de quarks samenbindt in hadronen, wordt overgedragen door gluonen g.

De consistentie van het SM vereist massaloze ijkbosonen, een gegeven dat duidelijk niet strookt met experimentele waarnemingen. Het Brout-Englert-Higgs mechanisme schiet hier te hulp om dit op te kunnen lossen: het Higgsveld krijgt spontaan een verwachtingswaarde voor het vacuüm welke ongelijk is aan nul, waardoor de oorspronkelijke elektrozwakke symmetrie breekt en de massa's van de ijkbosonen tot stand kunnen komen door de interacties tussen de ijkbosonen en het Higgsveld. Hetzelfde mechanisme geeft massa aan fermionen via zogenaamde Yukawa interacties.

Het Brout-Englert-Higgs mechanisme voorspelt het bestaan van een deeltje met onbekende massa: het welbekende Higgsboson. Een deeltje met een massa rond de 125 GeV, sterk gelijkend op het Higgs boson, werd in 2012 ontdekt door de ATLAS en CMS Collaboraties waarmee

de juistheid van het mechanisme experimenteel is aangetoond.

Wanneer de massa eenmaal vastgesteld is kunnen alle eigenschappen, zoals de koppeling met andere deeltjes, voorspeld worden. Tot op heden zijn er geen significante afwijkingen van de voorspelde waarden gevonden. Hoewel het SM een zeer succesvolle theorie blijkt, is het geen ultieme theorie van de natuur, daar het veel fundamentele vragen onbeantwoord laat zoals de aard van Donkere Materie en het "hiërarchie vraagstuk". Veel alternatieve theorieën die proberen deze vraagstukken op te lossen, voorspellen afwijkende eigenschappen voor het Higgsboson (of het bestaan van meerdere Higgsbosonen). Anderzijds, het gegeven dat er geen tekenen van nieuwe fysica zichtbaar zijn aan de horizon, maakt precieze metingen in de Higgs-sector een hoge prioriteit binnen de deeltjesfysica, om zo de interne consistentie van het SM te testen.

In dit opzicht is het meten van de Higgs-koppelingen aan de laatste generatie quarks van groot belang. Indirecte metingen van de top Yukawa-koppeling zijn uitgevoerd door de waarde af te leiden van het gluonfusieproces. Echter kunnen ook andere deeltjes bijdragen in dergelijke lusdiagrammen (zelfs hypothetische nieuwe deeltjes niet beschreven in het SM). Daartegenover is het mogelijk om deze koppeling direct te meten via het $t\bar{t}H$-productieproces.

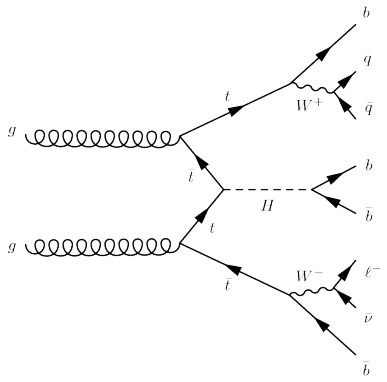


Figure S.1: Feynmandiagram van het $t\bar{t}H(b\bar{b})$-proces.

Dit proefschrift presenteert de zoektocht naar het standaardmodel Higgsboson dat geproduceerd wordt in combinatie met een topquarkpaar en vervalt in een $b\bar{b}$-paar, $t\bar{t}H(b\bar{b})$. Het topquark vervalt (bijna) uitsluitend in een W-boson en een b-quark. De topologie van het $t\bar{t}$-verval bepaalt de effectief gebruikte analysekanalen: dilepton en (gescheiden en "boosted") enkellepton. Het proces is weergegeven in een Feynmandiagram in Figuur S.1.

Deze analyse presenteert verscheidene problemen, allemaal herleidbaar tot de grote irreducibele $t\bar{t}$ + jets achtergrond. Met name in de

gekozen faseruimte van de analyse zijn de extra jets die worden geproduceerd in combinatie met het $t\bar{t}$-paar, vaak jets met een "zware smaak", d.w.z. jets afkomstig van b- of c-quark fragmentatie. Deze processen zijn niet alleen lastig te modelleren maar introduceren ook grote theoretische onzekerheden.

De serieuze problemen die de overweldigende $t\bar{t} + \geq 1b$-achtergrond met zich meebrengt, vereisen een complexe analysestrategie waarin verschillende multivariate technieken worden toegepast. De geselecteerde events worden eerst gecategoriseerd in signaal- en achtergrondverrijkte regio's. In de signaalregio's worden verschillende methoden gebruikt om de eindtoestand te reconstrueren; vervolgens worden Versterkte Beslissingsbomen (VBB) gebruikt om de events te classificeren in signaal- en achtergrondachtige events. Ten slotte, worden alle regio's gebruikt als input voor een (geprofileerde) meest-aannemelijkeschatting gegeven de data, om zo tegelijkertijd het aantal events voor de bestudeerde fysische processen te bepalen en om het algemene achtergrondmodel te beperken binnen de toegewezen systematische onzekerheden.

De grote combinatoriek als gevolg van de vele jets in de eindtoestand compliceert de identificatie van de vervalproducten van het Higgsboson en de reconstructie van de Higgs massapiek. De Jet Vertex Lading labeller (JVL) was specifiek ontwikkeld om dit aspect te verbeteren.

Het berust op het feit dat de jet-quark-toewijzing kan profiteren van de meting van de elektrische lading van het b-quark dat ten grondslag ligt aan de jetformatie, wat zowel de reconstructie van de eindtoestand als de identificatie van de Higgs-vervalproducten vergemakkelijkt en derhalve de gevoeligheid van de analyse vergroot. Zowel het algoritme als de kalibratie-analyse staan beschreven in Hoofdstuk 4.

JVL gebruikt de onderscheidende tekenen in de b-hadron-vervalketen om verschillende variabelen te construeren welke gevoelig zijn voor de quarklading en worden gecombineerd door middel van een neuraal netwerk. Uiteindelijk, wordt het aannemelijkheidsquotiënt van de hypothesen voor een b-jet om een positieve of negatieve lading te hebben gebruikt als uiteindelijke discriminant, λ_{JVC}, waarvan de verdeling is weergegeven in Figuur S.2. Daarmee is de JVL in staat om jets te labellen als afkomstig van een positief of negatief geladen b-hadron.

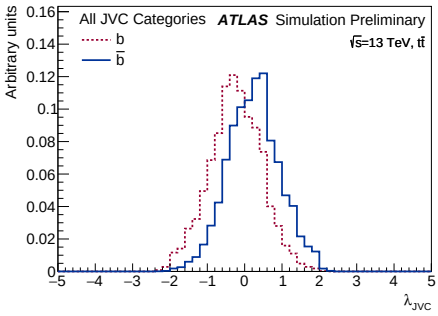


Figure S.2: Genormaliseerde λ_{JVC} verdeling voor positieve (ononderbroken blauwe lijn) en negatieve (gestippelde rode lijn) ware b-jets.

De Jet Vertex Lading-implementatie in de $t\bar{t}H(b\bar{b})$-analyse is beschreven in het eerste deel van Hoofdstuk 5, de ICHEP-analyse. Ondanks de verbetering in de reconstructie van de eindtoestand is er slechts een marginale verbetering in het onderscheiden tussen het $t\bar{t}H(b\bar{b})$-signaal en de dominante $t\bar{t} + \geq 1b$ achtergrond waargenomen, vanwege het feit dat de $t\bar{t} + \geq 1b$ achtergrond dezelfde ladingsignatuur heeft als het $t\bar{t}H(b\bar{b})$ signaal. Om deze reden is het niet gebruikt in de uiteindelijke analyse.

Om de kennis van de $t\bar{t} + \geq 1b$ en $t\bar{t} + \geq 1c$ achtergronden te verbeteren is de HFBDT ontwikkeld voor de ICHEP-analyse. Een VBB is getraind om de zware smaken van de $t\bar{t}$ +lichte smaak component te onderscheiden, zodat de (geprofileerde) meest-aannemelijke-schatting de bijbehorende systematische onzekerheden beter behandelt. Er werd echter geen significante verbetering waargenomen, dus is het niet gebruikt in het ontwerp van de uiteindelijke analyse.

Het tweede deel van Hoofdstuk 5 bevat de beschrijving van de zogenaamde “publicatie” analyse. Met betrekking tot de ICHEP-analyse is het een bijgewerkte versie die verschillende verfijningen en verbeteringen bevat, met name een verfijnde definitie van de signaal- en achtergrondregio’s. De resultaten zijn gebaseerd op data verkregen door $\sqrt{s}$ = 13 TeV proton-proton botsingen, met een geïntegreerde luminositeit van 36.1 fb^{-1} en verzameld met de ATLAS detector gedurende 2015 en 2016.

De best passende waarde voor de verhouding tussen de gemeten en voorspelde botsingsdoorsnede, ook wel de signaalsterkte μ genoemd, in alle enkel-lepton en dilepton regio’s levert een waarde op van:

$$\mu = \frac{\sigma_{\text{meas}}}{\sigma_{\text{SM}}} = 0.84 \pm 0.29\ (\text{stat.})\ ^{+0.57}_{-0.54}\ (\text{syst.}) = 0.84^{+0.64}_{-0.61}$$

met de verwachte onzekerheid op de signaalsterkte identiek aan de gemeten sterkte. Het komt overeen met een overschot van 1.4 standaarddeviaties boven de verwachte achtergrond van het SM.

Aangezien de $t\bar{t}H$ botsingsdoorsnede ongeveer 1% van de totale botsingsdoorsnede van het Higgs-deeltje uitmaakt zijn complementaire analyses in andere vervalkanalen uitgevoerd om zo extra gevoeligheid voor het $t\bar{t}H$-productieproces te verkrijgen: het Higgsboson verval naar een paar fotonen, $t\bar{t}H(\to \gamma\gamma)$; het Higgsboson verval in vier leptonen, $t\bar{t}H(\to ZZ^* \to 4\ell)$; en het verval in een multi-lepton eindtoestand, $t\bar{t}H(\to WW^*/\tau\tau/ZZ^* \to \text{leptons})$, aangeduid als $H \to \text{ML}$.

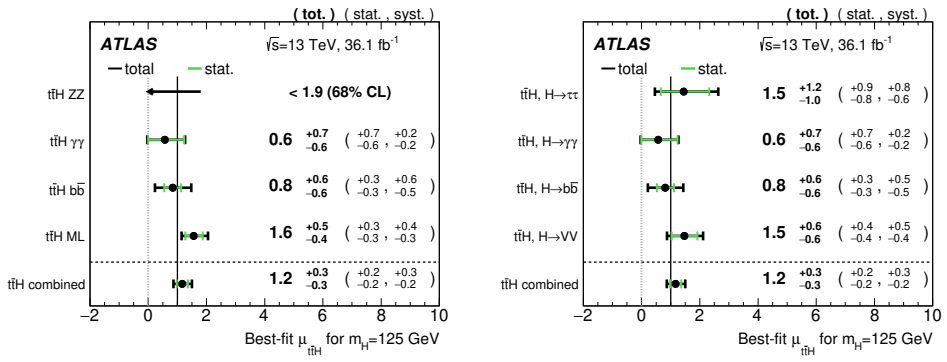
Deze analyses zijn ook uitgevoerd met een geïntegreerde luminositeit van 36.1 fb^{-1}. Een gecombineerde aannemelijkheidsfunctie die het mogelijk maakt om op de juiste manier rekening te houden met gemeenschappelijk systematische onzekerheden, wordt verkregen uit het product van de individuele aannemelijkheidsfuncties van elke analyse. De best passende waarde van de $t\bar{t}H$-signaalsterkte bepaald op basis van de gecombineerde aannemelijkheidsfunctie is gelijk aan:

$$\mu = 1.17 \pm 0.19\ (\text{stat.})\ ^{+0.27}_{-0.23}\ (\text{syst.})$$

wat overeenkomt met een overschot aan events boven de verwachte SM-achtergrond met een waargenomen (verwachte) significantie van 4.2 (3.8) σ: het is het eerste bewijs voor het $t\bar{t}H$-productiemechanisme dat wordt gevonden door ATLAS alleen. De botsingsdoorsnede voor de productie van $t\bar{t}H$ corresponderend met de best passende waarde van μ is 590^{+160}_{-150} fb, ruimschoots in overeenstemming met de SM-voorspelling van $\sigma^{\text{SM}}_{t\bar{t}H} = 507^{+35}_{-50}$ fb.

De waargenomen signaalsterkten voor de afzonderlijke analyses en hun combinatie zijn weergegeven in Figuur S.3a. Aangezien de vier analyses verschillende acceptaties hebben in hun analyseregio's voor de verschillende vervalkanalen van het Higgsboson is het mogelijk om μ onafhankelijk te bepalen voor verschillende Higgsvervallen. Figuur S.3b toont het resultaat van een fit aan vier signaalsterkten, één voor elk van

de vervallen $H \to \tau\tau$, $H \to \gamma\gamma$, $H \to b\bar{b}$ en $H \to WW/ZZ$. Alle metingen zijn in overeenstemming met de SM-voorspellingen en er is geen significante afwijking waargenomen.



(a) **(b)**

Figure S.3: Samenvatting van de metingen van μ uit individuele analyses en hun combinatie (links) en samenvatting van de best passende waarden voor μ uitgesplitst per vervalkanaal van het Higgsboson (rechts). Aangezien er geen events zijn waargenomen in de $H \to 4\ell$-analyse wordt er een bovengrens op μ gerapporteerd met een 68% betrouwbaarheidsniveau. De categorie $H \to VV$ combineert zowel $H \to WW^*$ als $H \to ZZ^*$.

About the author

Luca Colasurdo was born on 11 March 1989 in Campobasso, Italy. After graduating at the "Liceo Scientifico A.Romita" high-school, he enrolled at the University of Rome "Sapienza", where he obtained the bachelor's degree in October 2011.

In 2011, the author started his Master's at the "Sapienza" University and obtained his degree in November 2013. The Master's thesis is based on his contributions in the analysis searching for the Higgs boson in the decay channel into two tau leptons.

Throughout his stay in Rome, the author was based in the boarding school "Collegio Lamaro-Pozzani". During summer 2012 he participated in the CERN Summer Student Programme and from October 2012 to March 2013 he joined the Erasmus programme with destination Valencia.

In January 2014, the author started his PhD at the Radboud University in Nijmegen under the supervision of Dr. Frank Filthaut and Prof. Nicolo de Groot. His focus was first on the development of the Jet Vertex Charge algorithm and later in the search for the Higgs boson produced in association with a top quark pair and decaying into *b*-quarks, the results of which are presented in this thesis. He also took responsibilities as a teaching assistant at the Radboud University for particle Physics courses and a second-year laboratory class.

Acknowledgments

The moment of writing these pages has finally arrived, after what has been a long (more than it should have been) journey. It's not easy to thank properly everyone who contributed to this work and helped me during the past years in just a few line, but I'll try my best.

Firstly, I would like to thank my supervisors, Prof. Nicolo de Groot and Dr. Frank Filthaut, for believing in me and giving me this opportunity to grow and develop. Without your guidance I would have not been able to make it this far. I also want to thank the members of the manuscript committee, Prof. Frank Verbunt, Prof. Ronald Kleiss, Prof. Frank Linde, Prof. Bob van Eijk and Dr. Charles Timmermans, for carefully reading and reviewing this thesis.

A special thank you goes to Snežana, the best post-doc one could ever hope for. With your enthusiasm you taught me a lot, both personally and professionally; it was a pleasure to work with you.

A big thanks to the best secretaries I have ever met: Annelies, Marjo, Gemma and Gertie. You have always been kind, available and incredibly helpful in fighting bureaucracy every time I needed.

If these years have been full of laughter and nice memories, the credits go to my friends and colleagues: Alex, Alice, Anamika, Antonia, Antonio, Cristina, Daniil, Dominik, Evelin, Fabrizia, Giuseppe, Geert-Jan, Guus, Ivan, Jeroen, Lydia, Lucia, Lucrezia, Marie, Matteo, Nadezhda, Remco, Ruchi, Sara, Stefan, Tim, Veronica, Vince and all of the amazing people I met in these years. I do hope our paths will cross again in the future.

A special mention goes to my paranymphs, Giuseppe and Jeroen, to the Italians in Nijmegen and to Remco: thank you!

My gratitude goes also to my flatmates in Ferney-Voltaire, Fabrizio and Danilo, for making me feel at home instantly, for the nice dinners and all the fun we had.

A big, big thank you to Pina for the patience you took in designing this wonderful cover; and to all of my friends from Morrone: even if it's

only once a year, the calm, easygoing and sunny memories together are worth the wait.

Finally, mum and dad, thanks for having always given me the freedom to choose my own path, even if you were dubious at times about it. And thanks to my whole family: you don't choose it when you are born, but I consider myself lucky to be part of this one.

This line is not only the end of this work, but is the end of this chapter of my life as well. Thank you all for being a part of it.